Predicting sequence variant deleteriousness in genomes of livestock species

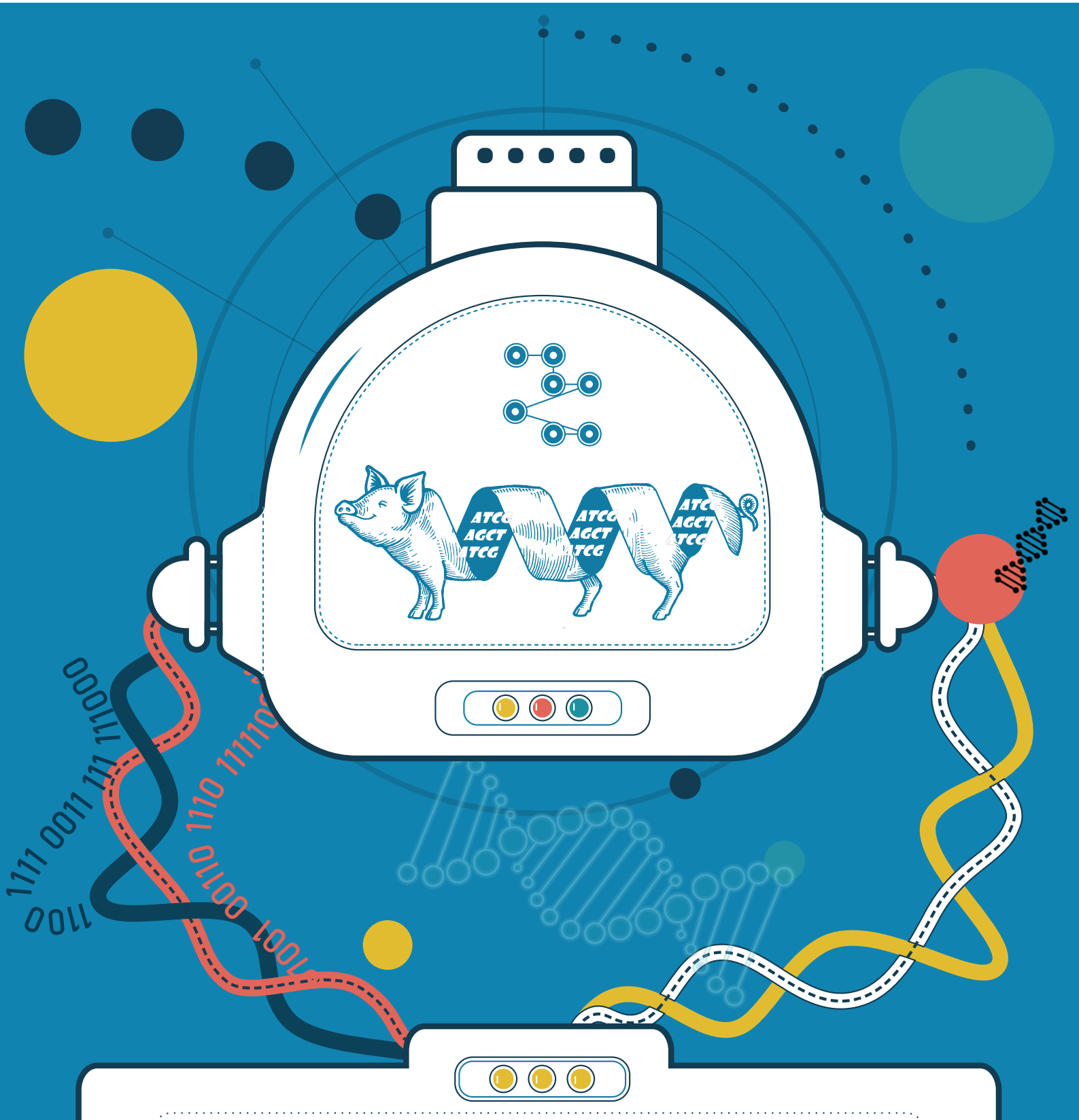
Groß, C.

DOI
10.4233/uuid:c9020486-82a6-4e4a-a9f5-f2b00ebc432c

Publication date
2020

Document Version
Final published version

Citation (APA)
Groß, C. (2020). *Predicting sequence variant deleteriousness in genomes of livestock species*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:c9020486-82a6-4e4a-a9f5-f2b00ebc432c



PREDICTING

SEQUENCE VARIANT DELETERIOUSNESS IN GENOMES OF LIVESTOCK SPECIES

CHRISTIAN GROß

Predicting sequence variant deleteriousness in genomes of livestock species

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Tuesday 13 October 2020 at 12:30 o'clock

by

Christian GROß
Master of Science in Bioinformatics
Vrije Universiteit Amsterdam, The Netherlands
born in Höxter, Germany

This dissertation has been approved by the

Promotor: Prof. dr. ir. M.J.T. Reinders and
Promotor: Prof. dr. ir. D. de Ridder

Composition of the doctoral committee:

Rector Magnificus,	chairman
Prof. dr. ir. M.J.T. Reinders	Delft University of Technology, promotor
Prof. dr. ir. D. de Ridder	Wageningen University and Research, promotor

Independent members:

Prof. dr. R.C.H.J van Ham	Delft University of Technology
Prof. dr. agr. habil. Dipl. Ing. agr. H. Simianer	University Goettingen, Germany
Dr. ir. P.D. Moerland	Amsterdam University Medical Center
Dr. C.F.H.A. Gilissen	Radboud University Medical Center

Other members:

Prof. dr. M.A.M. Groenen	Wageningen University & Research

Keywords: Functional genomics, DNA Variant Effect Prediction
Printed by: …
Front & Back: 路璐 (Lu Lu)

ISBN 000-00-0000-000-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/

Table of Contents

Summary

Illuminating the functional part of the genome of livestock species has the potential to facilitate precision breeding and to accelerate improvements. Identifying functional and potentially deleterious mutations can provide breeders with crucial information to tackle inbreeding depression or to increase the overall health of their populations and animal welfare. By performing Genome Wide Association Studies (GWAS) the genome can be interrogated for mutations that co-occur with a phenotype of interest. However, every GWAS delivers a large number of potentially functionally important single nucleotide polymorphisms (SNPs). The exact effect of each of these SNPs is often not known, especially for SNPs in noncoding sequences. Investigating each candidate SNP variant in detail is laborious and, eventually, infeasible, given the sheer number of variants. Thus, there is a strong need for approaches to select the most promising SNP candidates. Prioritizing variants, in particular, SNPs, has seen major developments in recent years which led to several discoveries and insights inheritable diseases of humans. Despite their great economical value, for livestock and other non-human species, this development is lagging behind.

A major contributing factor to the deficit in prioritization tools for non-human species is a lack of genomic annotations. In this thesis, we translated one of the currently popular SNP prioritization tools, CADD (Combined Annotation-Dependent Depletion), to mouse (mCADD) and performed an experiment in which we simulated a decrease in the number of available genomic annotations. These results showed that following the CADD approach to predict the putative deleteriousness of SNPs is meaningful in a non-human species, even when fewer genomic annotations are available than for the human case. This motivated us to build various CADD-like SNP prioritization tools for livestock species, in particular for pig (pCADD) and chicken (chCADD). We validated the pig prioritization tool on a set of well-known functional pig variants. Further, we showed how functional and non-functional parts of the pig genome are scored differently by pCADD. In collaboration with the breeding industry, we built upon the pCADD scores and implemented them in a pipeline to identify likely causal variants in GWAS. To this end, we utilized SNPs that were found significant in GWAS based on SNP-array data and found variants with high pCADD scores in whole genome sequence data that are in linkage disequilibrium with high GWAS-scoring SNPs. Thus, these pCADD-identified SNPs are likely (causal) functional candidates for the phenotypes tested. We also identified several expression quantitative loci (eQTL) variants, SNPs that explain observed differences in gene expression, which we were able to validate using RNA-seq data. This demonstrated the power of this new tool and its usefulness in identifying novel, functional variants. For chicken, we used the chCADD to interrogate highly conserved elements in the chicken genome. Here we found that, despite being highly conserved, not all parts of these elements might be functionally active. chCADD differentiates between regions within each conserved element that are predicted to be functionally different. Taken together, the results presented in this thesis demonstrate SNP prioritization can successfully be done in non-human species, which can greatly assist breeders and animal geneticists in their work to illuminate the functional genome.

1. Introduction

1.1. Effects of in-silico genome science on animal breeding

Humans have domesticated animals for around 12000 years [1] for the purpose of food production (e.g. livestock), protection (e.g. guard dogs), pest control (e.g. cats) and other functions. With these goals in mind, humans selected in particular the offspring of animals which displayed conducive traits to enhance these in future populations. In this way domestication differs from taming of animals, in which humans do not control the selection of mates to produce subsequent generations. By amplifying desirable traits in each generation, they become predominant in the controlled populations until the domesticated animals are clearly distinguishable from their wild counterparts, with their own characteristics.

For millennia, desirable traits have been selected based on visual inspection and evaluation of the mating candidates and their pedigrees. Statistical models were developed to predict breeding outcome (estimated breeding value (EBV) [2]) between two animals, to properly select the parent animals that have the highest potential to give birth to a generation of animals with improved phenotypes. Through this, animal breeding became more and more a theoretical subject in the natural sciences, with the constant goal of generating more accurate EBVs, to better select parent animals. One of the most widely used statistical models to calculate EBVs is the so-called best unbiased predictor (BLUP) [3]. It utilizes phenotype information and family relationships to calculate weighted phenotype averages that are corrected for potential systematic biases. Such biases include e.g. variation between farms, when differences in phenotypes are not due to differences in genetic value but differences in feed etc.

Breeders can use these predictions to formulate a breeding plan, which optimises the development of a trait in their populations. Through improvements in genomics, the development of genetic selection (GS) [4] and the ever decreasing costs of genome-wide single nucleotide polymorphism arrays (SNP arrays), this approach has been drastically enhanced in recent history. GS suggests the use of genetic markers rather than pedigree to identify the relatedness between individuals. This yields more accurate relationship estimates and more accurate EBVs. This progress helped to achieve major genetic improvements. From 1961 to 2008, egg, milk and meat production of major livestock species have increased by 20-30% due to improvements in genetics and other factors [5]. Broiler chickens in particular grew around three times faster in 2001 than in 1957, while consuming only a third of the feed [6].

Despite these improvements and the continuously growing amount of genomic data, the exact expression of a trait in any individual remains difficult to predict. This can be due to non-additive inheritance [7], underestimation of environmental effects that cause variations in the phenotype or absence of predictive genetic markers [8]–[10].

For genomic prediction, usually many tens or hundreds of thousands of SNPs measured in high throughput (genome wide association study (GWAS)) are considered as genetic markers. In these studies, associations between the genome and a phenotype of interest are usually found by analysing the overrepresentation of SNPs between two different cohorts of individuals. Livestock species usually carry 2 to 4 times more mutations than humans [11]–[14]. Combined with the fact that 25-50% of rare non-synonymous mutations in humans are predicted to have an adverse effect on the survivability of the individual [15], a relatively larger number of genomic variants with adverse effects on survivability or phenotype can be assumed to be present in any genome of livestock species. These mutations, especially heterozygous occurrences of homozygous recessive variants, can stay present in the population at low frequencies. Due to high rates of inbreeding in breeding plans to emphasize the expression of a particular trait, even low frequency, heterozygous variants can frequently become homozygous and have adverse effects on the phenotypes. They

jointly lower the performance and fitness of the population, but due to their low frequency they are hard to identify and remove.

GWAS are the go-to approach to investigate associations between the genotype and phenotype. Even though at first view the approach to look for over-/under representations of alleles in two cohorts seems relatively straight forward, it is based on a number of assumptions and may be difficult to conduct without introducing biases. To find all variants in a GWAS that have an effect on the phenotype, all variations of all individuals need to be tested. In humans, persons differ by around 8-10 million SNPs from each other and around 40 million base pairs are affected by structural variations [16], which may differ between cohorts as well. This means the number of tested people need to be enormous, otherwise there is no chance that any variation may be identified due to low statistical power. In animal GWAS, the same problem of low statistical power occurs. To increase power, either more samples can be used or fewer alleles. Most often the number of samples cannot be increased, thus fewer alleles are chosen. A selected number of marker SNPs can still be informative about the genotype because SNPs in close proximity are often inherited together, so a SNP can give information (SNP imputation) [17] on nearby SNPs even if they are not measured directly. The mutual inheritance of SNPs is called linkage disequilibrium [18] (LD). The sizes of these LD-blocks, which are inherited together, depend on the degree of inbreeding, with more inbreeding leading to larger LD-blocks. For this reason, marker SNPs should be carefully selected for GWAS to represent LD blocks associated with the phenotype of interest [19]–[21]. Still, errors may accumulate and the observed change of phenotype, caused by each marker, does not necessarily sum up to change that would be expected [22], [23].

SNPs constitute the most common and most easily measured type of genetic variation, hence the strong emphasis on these in GWAS. In GWAS, SNPs located in the same LD-block are highly scored if that LD-block segregates between the two tested cohorts of animals. Generally, it is assumed that there is only one variant per highly scored LD-block which affects the investigated phenotype, while the other marker SNPs are only linked to that causal/functional mutation through LD. Due to the low likelihood of truly causal SNPs being selected in the subset of measured SNPs, the results of GWAS have to be further scrutinized. LD-blocks differ in size and can range over several millions of base pairs (Mbp), covering numerous genes. As manual identification of the causal variant is infeasible, *in silico* SNP prioritization tools have been developed. These tools often calculate a specific metric for each SNP; one of these is its expected deleteriousness. Deleteriousness is not clearly defined and can have several meanings. First there is the gene-centric definition of deleteriousness. It states that the SNP has an adverse effect on a gene, either by lowering its expression or disrupting the structure of the encoded protein, rendering it incapable of performing its function. Another definition is centered around evolution, identifying deleteriousness as the likelihood of a SNP to be under negative selection due to a disadvantageous effect on the phenotype that decreases the probability of the individual to reproduce.

Besides the use of SNP prioritisation tools in GWAS, they may be able to help to study functional elements across the genome which would eventually support selection in breeding plans. So far, in animal breeding, the genome has been used as a black box, emphasizing genomic loci rather than individual functional mutations. Illuminating the genomic black box could lead to improved weighting of SNPs in genomic breeding, which has the potential to greatly increase genetic gains in of the studied animals.

The identification of functional elements and SNPs depends on the region of the genome in which they are supposed to be located. Until recently this has been the reason why great emphasis was put on the identification of functional SNPs in exonic regions, while SNPs in other regions have been neglected due to the difficulty to infer causality for their function. In the past, this has led to the incorrect assumption that sequences which do not code for a gene were unimportant. This

assumption was coined in the term "junk DNA" [24]. Since then, many regulatory active regions that are essential for survival of the individual, have been identified within the non-coding part of the genome [25]–[28]. Thus, there is a great need to investigate variations in all parts of the genome. SNP prioritisation tools, capable of scoring variants genome-wide, may be able to help identify and discriminate functional from non-functional DNA sequences. In this way they provide an order of importance, to study variations that could complement and improve existing breeding methods.

1.2. Metrics and methods for SNP prioritization

To prioritize SNPs, we first have to identify properties of SNPs related to the evaluation metric used. A SNP in itself has only three distinct properties: its location on the genome, frequency within the population and the nucleotide substitution it represents. The most informative property is its location. When researchers investigate the effect of a particular SNP on a phenotype, they can derive conclusions based on additional annotations known at that location. The number and diversity of these additional annotations differ greatly between genomic regions and species. A SNP located in a known exon may have many more annotations than SNPs in other parts of the genome; moreover, by taking advantage of the nucleotide substitution which the SNP represents, potential effects downstream of protein production can be inferred. The first and most common kind of SNP effect prediction and prioritization tools are specific for these information-rich genomic positions. Tools such as SIFT [29], PolyPhen & PolyPhen-2 [30], [31], SNAP & SNAP2 [32], [33] and Provean [34] make use of amino acid conservation and the potential effect of an amino acid substitution on the function of the protein. Unfortunately, in mammals only roughly 1%-3% of the genome codes for protein [35], which limits the overall use of missense specific SNP prioritisation tools. Further, it has been estimated that non-synonymous mutations only account for 20% of the genetic variation that influences a change of phenotype [36]. The majority of the genome does not code for a protein and the majority of influential loci, identified in GWAS, are located in regions that are not annotated with any genes or that belong to the noncoding part of a gene. Regulatory elements which have an effect on gene expression and phenotype are often located in these regions. This is the main motivation behind the push to develop more elaborate SNP prioritization tools capable of annotating mutations in noncoding DNA sequences.

Due to the complexity of any trait, its expression depends on a plethora of interacting regulatory programs that work together and create the observable phenotype. At each stage, from DNA to RNA to protein, regulatory effects manifest themselves. These regulatory effects are caused either by cis- or trans-regulatory elements. Cis regulatory elements are located in the DNA sequence, most likely in close proximity to the regulated gene, i.e. promoter regions of a gene. While promotors are always close to their associated gene, enhancer and silencer regions may be more distant to their target and have to be identified via measurements of the quaternary structure of the DNA or other specific characteristics of those regions such as their methylation and acetylation status. Trans-regulatory elements are elements such as transcription factors or miRNAs which are not located on the same DNA molecule as the regulation target. Each of these elements can be measured in various ways, resulting in many different data types. SNP prioritisation tools usually capitalize on this data to prioritize variants in non-coding regions.

DNA quaternary structure is of importance because of the densely packed nature of DNA which only allows for the binding of transcription factors (TF) at exposed sites. These can be experimentally identified via FAIRE-seq, DNase-seq and ChiP-seq assays [37]–[39]. Due to the relative importance of these regulatory regions, several methods have used the rich data sets of the ENCODE data base [40] to learn predictors for sequence motifs which indicate DNA accessibility and TF binding sites. Three of these methods (DeepBind, DeepSea, Basset [41]–[43]) are suitable

to varying degrees to make predictions about the putative effect of SNPs in predicted regulatory active regions. DeepBind is the least optimized for the prioritization of SNP: based on ChiP-seq data, it predicts sequence motifs and scores sequences containing those. It then uses these sequence scores to score individual SNPs contained in them. DeepSea is more tailored to predict the functionality of SNPs at single bp resolution. It is a deep learning approach that uses a convolutional layer to learn features from the DNA sequence which are informative for DNase activity, TF binding and histone marks. Then it predicts how these features change when the investigated sequence harbours a SNP relative to the reference sequence. Basset takes a similar approach but predicts only DNA accessibility via DNase activity. It sets itself apart from the other two methods by predicting the change of DNA accessibility per cell type, which allows for the prioritization of SNPs with respect to cell type specific traits.

Disadvantages of all these methods are that they are limited to specific regions, similar to the previously mentioned missense specific methods. Missense specific methods are limited to amino acid changing mutations; DeepBind, DeepSea and Basset are limited to accessible noncoding regions but ignore SNPs in other regions. Further, epigenetics differs from cell type to cell type and can change with age and other environmental circumstances, thus SNPs functional for the trait of interest may not be detected in the investigated sample. It has been shown that in some cases, the epigenome influences gene expression more than sequence and that some epigenetic markers are more strongly conserved than the sequence in that region, which allows for the introduction of SNPs without major changes in gene expression [44]. This means SNP effects are hard to derive from genome sequence alone. Finally, the three methods rely to a large extent on the vast amount of data available in public databases for human genome research. For the purpose of creating a SNP prioritization tool for non-human species, these approaches are less suitable due to a lack of publicly available data. This is even true for model organisms. ENCODE v93 (accessed 06-12-2019) contains the results of 10,485 assays for human while for mouse there are only 1916; the database which is supposed to be established as part of the FAANG project [45], aiming to be the counterpart to ENCODE for livestock genomes, is at the beginning of December 2019 [46] still in its early development.

A similar problem is observed when one wants to recreate models for non-human species that are trained on known disease SNPs. Human examples of such models are FATHMM-MKL [47] or GWAVA [48]. Both use data sets of experimentally validated disease-associated variants for training. Such data sets are not available to the same extent for other species, which limits the portability of these methods. Moreover, problems may arise due to the variants used for training. It can be hypothesized that these represent an extreme subset of functional variants and therefore any model learned on them would have difficulty to differentiate well among less extreme variants.

SNP prioritization approaches such as Combined Annotation Dependent Depletion [49] (CADD) and linear-INSIGHT [50] (LINSIGHT) avoid these kinds of problems. Instead of training a model on a small set of already validated variants, they use evolutionary models to capture signals of natural selection over many generations. In this way, they obtain large numbers of variants that can be used to train models which emphasize the discrimination of variants under purifying selection. LINSIGHT uses the INSIGHT [51] evolutionary model that estimates which regions are under purifying selection by contrasting them to neutrally evolving regions. To do this, it uses differences between population and outgroup variants. Then a generalized linear model is trained to predict the INSIGHT classification based on genomic annotations such as conservation scores, TF binding sites and epigenetic markers. The resolution of the score can range from single bp to several kbp. CADD on the other hand does not make assumptions about entire regions of negative selection. It relies more strongly on the inference of past ancestral states, since it derives a nucleotide substitution model from substitutions between different ancestral genomes. This substitution model is then employed to simulate *de novo* variants which are more likely to have experienced negative

selection and therefore are enriched in deleterious variants. This set of simulated variants is used as a positive set and the negative set contains derived alleles which are (almost) fixed in the species of interest. These alleles have experienced many generations of selection pressure and should be depleted of variants with an adverse effect on the phenotype. As in the previously discussed methods, all variants are annotated with a wide range of genomic annotations to train a machine learning model to differentiate between both classes. In CADD this is a penalized linear logistic model. In comparison to LINSIGHT, CADD has a single nucleotide resolution genome-wide, with individual scores for different alleles at the same site, and it incorporates coding and noncoding regions while LINSIGHT is trained particularly for non-coding DNA.

1.3. Thesis outline / contributions

The research presented here focuses on the use of the CADD approach for non-human species. CADD is based on a single model for the entire genome and has been well received in the investigation of human genomes [52]–[56]. Its general framework can be reproduced for any species as long as whole genome sequences of at least three other closely related species are known. First, Chapter 2 presents a feasibility study in mouse, demonstrating that CADD methodology can be meaningfully reproduced for other non-human species, even when fewer genomic annotations are available. Chapter 3 follows up on these results and introduces the CADD methodology for pig. Chapter 4 shows the insights and value which can be generated by incorporating pig-CADD (pCADD) in the prioritisation process of SNPs in the breeding environment. Finally, Chapter 5 introduces chicken-CADD (chCADD) and exploits its single allele resolution to investigate highly conserved regions in chicken, for which detailed genomic annotations are missing.

Bibliography

[1] F. Luca, G. H. Perry, and A. Di Rienzo, "Evolutionary adaptations to dietary changes," *Annu. Rev. Nutr.*, vol. 30, no. 1, pp. 291–314, 2010.

[2] P. VanRaden *et al.*, "Invited review : Reliability of genomic predictions for North American Holstein bulls," *J. Dairy Sci.*, vol. 92, no. 1, pp. 16–24, 2009.

[3] C. R. Henderson, *Applications of linear models in animal breeding models*. University of Guelph, 1984.

[4] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[5] P. K. Thornton, "Livestock production: recent trends, future prospects," *Philos. Trans. R. Soc.*, vol. 365, pp. 2853–2867, 2010.

[6] G. B. Havenstein, P. R. Ferket, and M. A. Qureshi, "Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets," *Poult. Sci.*, vol. 82, pp. 1500–1508, 2003.

[7] D. K. Seymour *et al.*, "Genetic architecture of nonadditive inheritance in Arabidopsis thaliana hybrids," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 1, pp. E7317–E7326, 2016.

[8] J. Felson, "What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption," *Soc. Sci. Res.*, vol. 43, pp. 184–199, 2014.

[9] F. Clerget-Darpoux and R. C. Elston, "Will formal genetics become dispensable ?," *Hum. Hered.*, vol. 76, pp. 47–52, 2013.

[10] S. J. Schrodi *et al.*, "Genetic-based prediction of disease traits: prediction is very difficult, especially about the future," *Front. Genet.*, vol. 5, pp. 1–18, 2014.

[11] M. A. M. Groenen *et al.*, "Analyses of pig genomes provide insight into porcine demography and evolution," *Nature*, vol. 491, no. 7424, pp. 393–398, 2012.

[12] M. Bosse *et al.*, "Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression," *Nat. Commun.*, vol. 5, pp. 1–8, 2014.

[13] G. K. Wong *et al.*, "A genetic variation map for chicken with polymorphisms," *Nature*, vol. 432, pp. 717–722, 2005.

[14] H. D. Daetwyler *et al.*, "Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle," *Nat. Genet.*, vol. 46, no. 8, pp. 858–865, 2014.

[15] D. M. Altshuler *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[16] The 1000 Genomes. Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, 2015.

[17] M. Stephens and P. Donnelly, "Inference in molecular population genetics," *J. R. Stat. Soc.*, vol. 62, no. 4, pp. 605–655, 2000.

[18] M. Slatkin, "Linkage disequilibrium: Understanding the genetic past and mapping the medical future," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–485, 2008.

[19] B. V Halldorsson *et al.*, "Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies," *Genome Res.*, vol. 14, pp. 1633–1640, 2004.

[20] Z. Meng, D. V Zaykin, C.-F. Xu, M. Wagner, and M. G. Ehm, "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes," *Am. J. Hum. Genet.*, vol. 73, pp. 115–130, 2003.

[21] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *Am. J. Hum. Genet.*, no. 74, pp. 106–120, 2004.

[22] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 4, pp. 1193–1198, 2012.

[23] S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher, "Estimating missing heritability for disease from genome-wide association studies," *Am. J. Hum. Genet.*, vol. 88, no. 3, pp. 294–305, 2011.

[24] C. F. Ehret and G. De Haller, "Origin, development and maturation of organelles and organelle systems," *J. Ultrastruct. Res.*, vol. 23, 1963.

[25] B. S. Gloss and M. E. Dinger, "Realizing the significance of noncoding functionality in clinical genomics," *Exp. Mol. Med.*, vol. 50, no. 8, 2018.

[26] D. E. Dickel *et al.*, "Ultraconserved enhancers are required for normal development," *Cell*, vol. 172, no. 3, pp. 491-499.e15, 2018.

[27] I. Dunham *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[28] A. F. Palazzo and T. R. Gregory, "The case for Junk DNA," *PLoS Genet.*, vol. 10, no. 5, 2014.

[29] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Res.*, vol. 11, no. 5, pp. 863–874, 2001.

[30] V. Ramensky, "Human non-synonymous SNPs: server and survey," *Nucleic Acids Res.*, vol. 30, no. 17, pp. 3894–3900, 2002.

[31] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[32] Y. Bromberg and B. Rost, "SNAP: Predict effect of non-synonymous polymorphisms on function," *Nucleic Acids Res.*, vol. 35, no. 11, pp. 3823–3835, 2007.

[33] M. Hecht, Y. Bromberg, and B. Rost, "Better prediction of functional effects for sequence variants," *BMC Genomics*, vol. 16, no. 8, p. S1, 2015.
[34] Y. Choi and A. P. Chan, "PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics*, vol. 31, no. 16, pp. 2745–2747, 2015.
[35] G. E. Sims, S. R. Jun, G. A. Wu, and S. H. Kim, "Whole-genome phylogeny of mammals: Evolutionary information in genic and nongenic regions," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 40, pp. 17077–17082, 2009.
[36] C. P. Ponting and R. C. Hardison, "What fraction of the human genome is functional?," *Genome Res.*, vol. 21, no. 11, pp. 1769–1776, 2011.
[37] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin," *Genome Res.*, vol. 17, no. 6, pp. 877–885, 2007.
[38] A. P. Boyle *et al.*, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, no. 2, pp. 311–322, 2008.
[39] D. S. Johnson, A. Mortazavi, and R. M. Myers, "Genome-wide mapping of in vivo protein-DNA interactions," vol. 316, no. June, pp. 1497–1503, 2007.
[40] R. M. Myers *et al.*, "A user's guide to the Encyclopedia of DNA elements (ENCODE)," *PLoS Biol.*, vol. 9, no. 4, 2011.
[41] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.
[42] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model.," *Nat. Methods*, vol. 12, no. 10, pp. 931–4, 2015.
[43] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Res.*, vol. 26, no. 7, pp. 990–999, 2016.
[44] S. Xiao *et al.*, "Comparative epigenomic annotation of regulatory DNA," *Cell*, vol. 149, no. 6, pp. 1381–1392, 2012.
[45] L. Andersson *et al.*, "Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project," *Genome Biol.*, vol. 16, no. 1, pp. 4–9, 2015.
[46] "Fragencode." [Online]. Available: http://www.fragencode.org/results.html. [Accessed: 06-Dec-2019].
[47] H. A. Shihab *et al.*, "An integrative approach to predicting the functional effects of non-coding and coding sequence variation," *Bioinformatics*, vol. 31, no. 10, pp. 1536–1543, 2015.
[48] G. R. S. Ritchie, I. Dunham, E. Zeggini, and P. Flicek, "Functional annotation of noncoding sequence variants," *Nat. Methods*, vol. 11, no. 3, pp. 294–296, 2014.
[49] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: Predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, 2019.
[50] Y. F. Huang, B. Gulko, and A. Siepel, "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data," *Nat. Genet.*, vol. 49, no. 4, pp. 618–624, 2017.
[51] I. Gronau, L. Arbiza, J. Mohammed, and A. Siepel, "Inference of natural selection from interspersed genomic elements based on polymorphism and divergence," *Mol. Biol. Evol.*, vol. 30, no. 5, pp. 1159–1171, 2013.
[52] J. K. van der Velde *et al.*, "Evaluation of CADD Scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization," *Hum. Mutat.*, vol. 36, no. 7, pp. 712–719, 2015.
[53] M. Mesbah-Uddin, R. Elango, B. Banaganapalli, N. A. Shaik, and F. A. Al-Abbasi, "In-silico analysis of inflammatory bowel disease (IBD) GWAS loci to novel connections," *PLoS One*, vol. 10, no. 3, pp. 1–21, 2015.
[54] H. Holstege *et al.*, "Characterization of pathogenic SORL1 genetic variants for association with Alzheimer ' s disease : a clinical interpretation strategy," *Eur. J. Hum. Genet.*, vol. 25, no. April, pp. 973–981, 2017.
[55] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuman, "Functional mapping and annotation of genetic associations with FUMA," *Nat. Commun.*, vol. 8, no. 1826, pp. 1–10, 2017.
[56] B. Banaganapalli *et al.*, "Comprehensive computational analysis of GWAS loci identifies CCR2 as a candidate gene for Celiac disease pathogenesis," *J. Cell. Biochem.*, vol. 118, no. 8, pp. 2193–2207, 2017.

2. Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse

Christian Groß

Dick de Ridder

Marcel Reinders

This chapter is published as: C. Groß, D. de Ridder & M. Reinders "Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse." *BMC Bioinformatics* 19, 373 (2018). https://doi.org/10.1186/s12859-018-2337-5

2.1. Abstract

Background: Predicting the deleteriousness of observed genomic variants has taken a step forward with the introduction of the Combined Annotation Dependent Depletion (CADD) approach, which trains a classifier on the wealth of available human genomic information. This raises the question whether it can be done with less data for non-human species. Here, we investigate the prerequisites to construct a CADD-based model for a non-human species.

Results: Performance of the mouse model is competitive with that of the human CADD model and better than established methods like PhastCons conservation scores and SIFT. Like in the human case, performance varies for different genomic regions and is best for coding regions. We also show the benefits of generating a species-specific model over lifting variants to a different species or applying a generic model. With fewer genomic annotations, performance on the test set as well as on the three validation sets is still good.

Conclusions: It is feasible to construct species-specific CADD models even when annotations such as epigenetic markers are not available. The minimal requirement for these models is the availability of a set of genomes of closely related species that can be used to infer an ancestor genome and substitution rates for the data generation.

2.2. Background

With the possibility of determining variation in genomes at large scale came an interest in predicting the influence of a mutation on a phenotype, in particular its pathogenicity. Initially, such predictions were restricted to missense mutations, as these cause a change in the corresponding amino acid chains and are thus most likely to have immediate functional effects. SIFT [1], PolyPhen2 [2], SNAP2 [3] and Provean [4] are examples of this kind of predictor. Recently, a number of methods for variant annotation were proposed that assign a single deleteriousness score to mutations throughout the entire genome, based on a large collection of genomic and epigenomic measurements. These methods – a.o. CADD [5], GWAVA [6], FATHMM-MKL [7] – are based on supervised classification. CADD (Combined Annotation Dependent Depletion) takes an interesting approach, in that it trains classifiers to distinguish between observed benign variants and inferred, putatively deleterious variants, instead of exploiting only known regulatory or disease-associated variants. This opens up the possibility to reproduce this approach for other non-human species as well. It shares similarities with fitCons [8] and LINSIGHT [9] by exploiting evolutionary models, which capture signals of natural selection over many generations in the generation of training data.

Although the use of CADD is already well-established in human genetics research and clinical practice [10], [11], for non-human species the situation is quite different. While generic predictors such as SIFT, Provean and SNAP2 can be used, genome-wide variant annotation methods are generally not available. A major reason is that for non-human genomes fewer genomic annotations are available, complicating the construction of more advanced models. This is even the case for model organisms, such as zebrafish (Danio rerio), drosophila (Drosophila melanogaster) and mouse (Mus musculus). Additionally, extensive population studies similar to the 1,000 and 100,000 Genomes Projects [12], [13] are lacking for non-human species, hampering the creation of good training data sets. Finally, models for non-human species are much more difficult to evaluate due to a lack of known disease-associated or phenotype-altering variants such as ClinVar offers for human [14].

Here, we explore the development of a functional prioritization method for SNVs located across the entire genome of a non-human species. The species we selected to investigate is mouse. As a

model species it is well studied, with relatively rich, publicly available, genomic annotation data sets [15]–[20]. Even though not all annotations used in the human CADD model are available for mouse, the large overlap of annotations allows performance evaluation and comparison between the original CADD and our mouse CADD. With this proof-of-principle, we aim to gain insight into design choices for porting such a methodology to non-human species.

2.3. Results

We trained a CADD model on mouse data (mCADD) and a CADD model on human data (hCADD). Performances of both are evaluated on test sets of variants located in different genomic regions. In addition, mCADD is evaluated on three validation sets (Fairfield, Mutagenetix, ClinVar-ESP data sets). We also compared mCADD to benchmark metrics such as SIFT and two PhastCons scores based on two phylogenies of different depth. Further, we trained mCADD and hCADD on four different annotation subsets to investigate the performance of a CADD-like classifier for species with fewer known annotations. These models are referred to as hCADD(n) and mCADD(n), with n the number of annotations used during training. To investigate the benefits of developing species-specific CADD models, we compared mCADD to 1) CADD v.1.3. C-scores by lifting validation variants from mm10 to hg19 , and 2) a CADD model trained on human data which, without further adaptation, is applied on mouse data to evaluate the mouse SNVs (hCADD*).

2.3.1. mCADD performs similarly on mouse as hCADD does on human

The ROC-AUC performance of mCADD(931) on the entire test set equals 0.668 (Figure 1), which is similar to the performance of hCADD(1000) applied on human data (Figure 2). Overall, mCADD(931) has a better performance across all genomic regions, with the most pronounced difference for the translated missense variants. Both models, mCADD(931) and hCADD(1000), discriminate between simulated and derived better than SIFT and PhastCons scores.

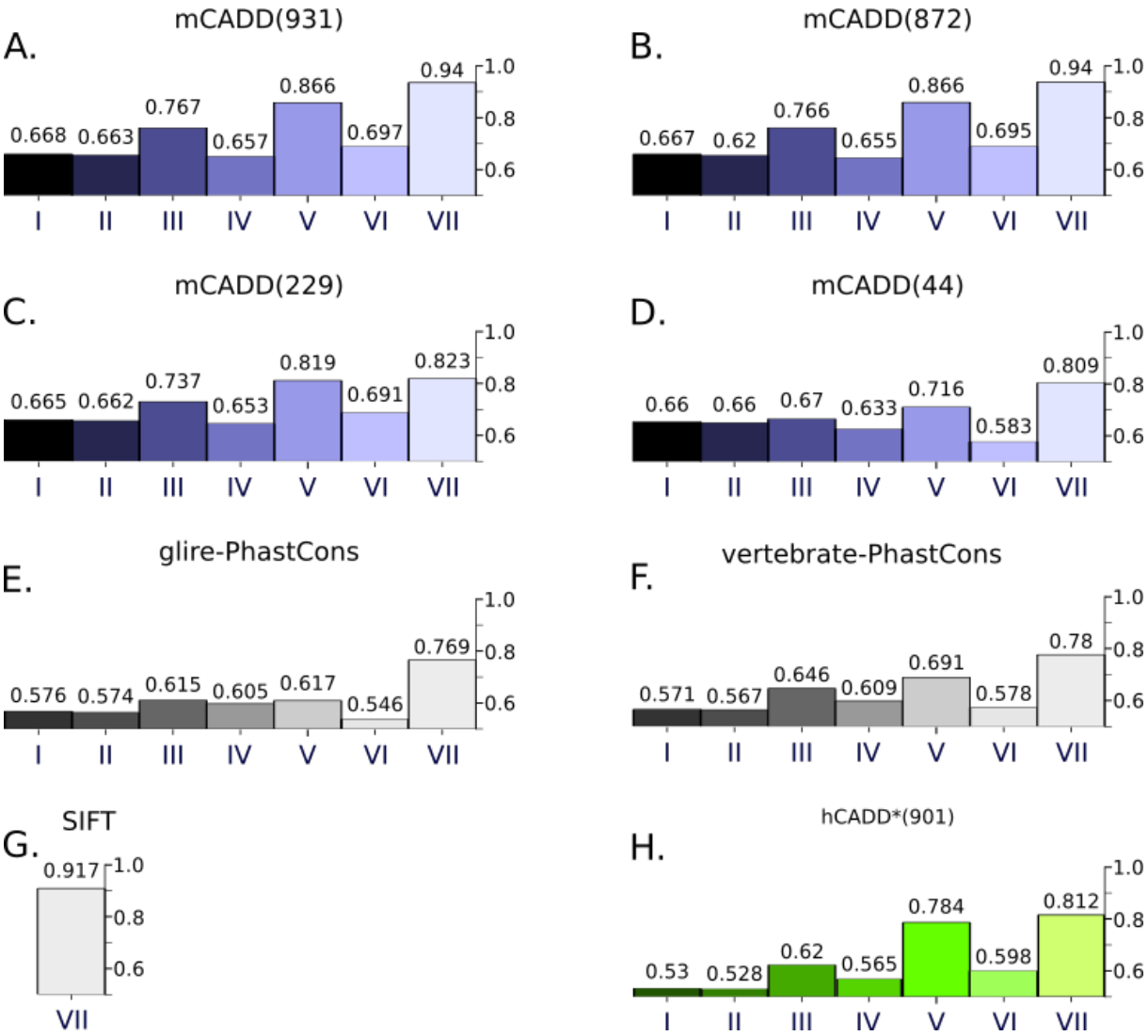


Figure 1: a-d) ROC-AUC scores of the four different mCADD models evaluated on seven different subsets of the mouse held-out test set reflecting different genomic regions and/or functional annotations. e, f) Seven different subsets of the mouse held-out test set evaluated by glire- and vertebrate based PhastCons scores, respectively. g) Missense mutations of the mouse held-out test set evaluated by SIFT. h) The subsets of the mouse held-out test set evaluated by hCADD*.: I) all data, II), not-transcribed, III) transcribed, IV) transcribed but not translated, V) translated, VI) translated and synonymous, and VII) translated and missense. The different models are indicated at the top of the panel. All displayed scores are ROC-AUC.

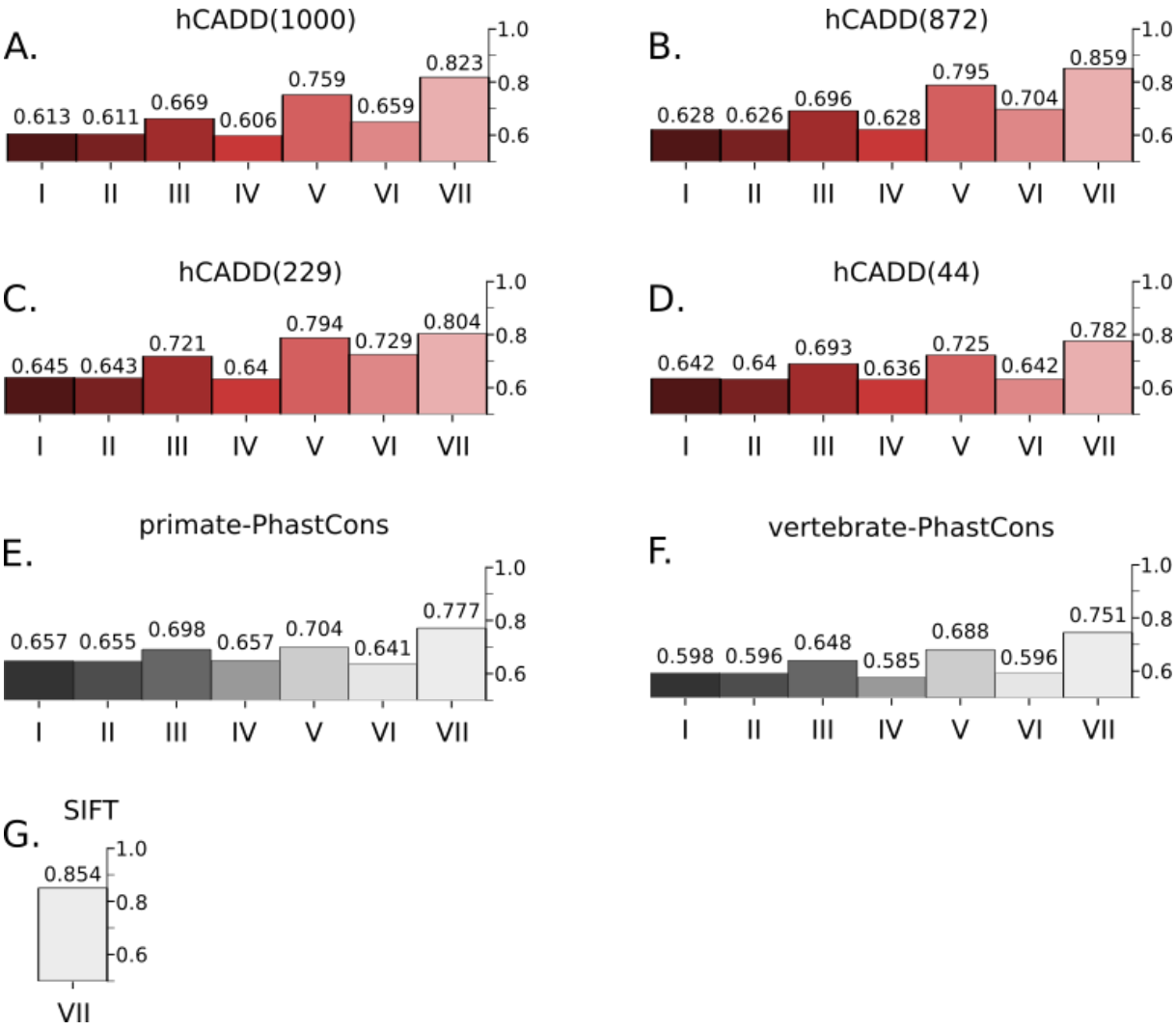


Figure 2: ROC-AUC scores of the four different hCADD models evaluated on the human held-out test set. e, f) Seven different subsets of the human held-out test set evaluated by primate- and vertebrate based PhastCons scores, respectively. g) Missense mutations of the human held-out test set evaluated by SIFT. (see caption Figure 1 for remaining explanation).

It is known that the distribution of CADD scores differs between genomic regions, and that the disruptive effect of variants in exonic regions can be estimated more precisely than that of variants in non-coding regions [21], [22] We observe a similar trend for mCADD(931) as well as hCADD(1000). Most of the performance increase from genomic regions I, III, V to VII (Figure 1) is even due to the high performance on correctly classifying missense mutations that become more enriched in these regions. This is in contrast to the performances in genomic regions II, IV and VI which do not contain any missense mutations.

2.3.2. Models trained on selected annotation subsets experience performance drop in coding Regions

To see whether models behave differently when less information is available, we reduced the number of annotations to train human and mouse models. The first subset of annotations (872) was chosen based on the idea that epigenetic measurements and species-specific annotations might not be available for some species. The performances of mCADD(931) and hCADD(1000) as

well as mCADD(872) and hCADD(872) are very similar, with the mCADD models performing slightly better than the hCADD models (Figure 1 and Figure 2).

The second subset of annotations consist of 229 annotations derived from sequence only, i.e. conservation scores and VEP consequences (mCADD(229), hCADD(229)). The situation is now different. The trend is still that performance increases from non-coding to coding to missense mutations. Also, SNVs in non-coding regions can still be classified with a performance comparable to that of models with more annotations. However, with the loss of particular information about coding regions and SIFT as an annotation, the performance of mCADD(229) to evaluate missense mutations drops below that of SIFT.

The smallest subset (44 annotations) excludes the VEP consequences and solely contains conservation scores and sequence features (mCADD(44), hCADD(44)). Now performances drop even further, but mCADD(44) shows that a simple combination of sequence based features and conservation scores outperforms the PhastCons scores for all genomic regions.

Interestingly, hCADD* (the human trained model applied on mouse data) performance lays between mCADD(229) and mCADD(44) for all translated regions (see Figure 1 V-VII) and is better than the PhastCons scores for those variant sets. On the other hand, hCADD* shows mostly random performance when non-translated regions are considered, indicating it is necessary to adapt the CADD model to species-specific data.

Taken together, decreasing the number of available annotations decreases performance, which drops relatively faster in coding regions than in non-coding regions. The drop in performance between mCADD(931) and mCADD(872) is, however, negligible, suggesting that epigenetic and species-specific annotations can be safely ignored.

2.3.3. Evaluation of phenotype affecting SNVs by mCADD

To show that mCADD is capable of accurately scoring real data and not only differentiates between simulated and derived variants, we evaluated the different mCADD models on three independent validation sets (see Figure 3). mCADD(931) and mCADD(872) perform extremely well on all three validation sets (ROC-AUC > 0.95) and hardly differ (see Figure 3). mCADD(229) performs comparably well on the ClinVar-ESP data set and shows a drop in performance on the Fairfield and Mutagenetix data sets. The drop increases when fewer annotations are considered for training (mCADD(44)). All mCADD models and hCADD* perform better than the two conservation scores, except for mCADD(44) on the Mutagenetix data. On all validation sets, the hCADD* performance lays between the performances of mCADD(229) and mCADD(44) and has relatively good performance on the ClinVar-ESP data set.

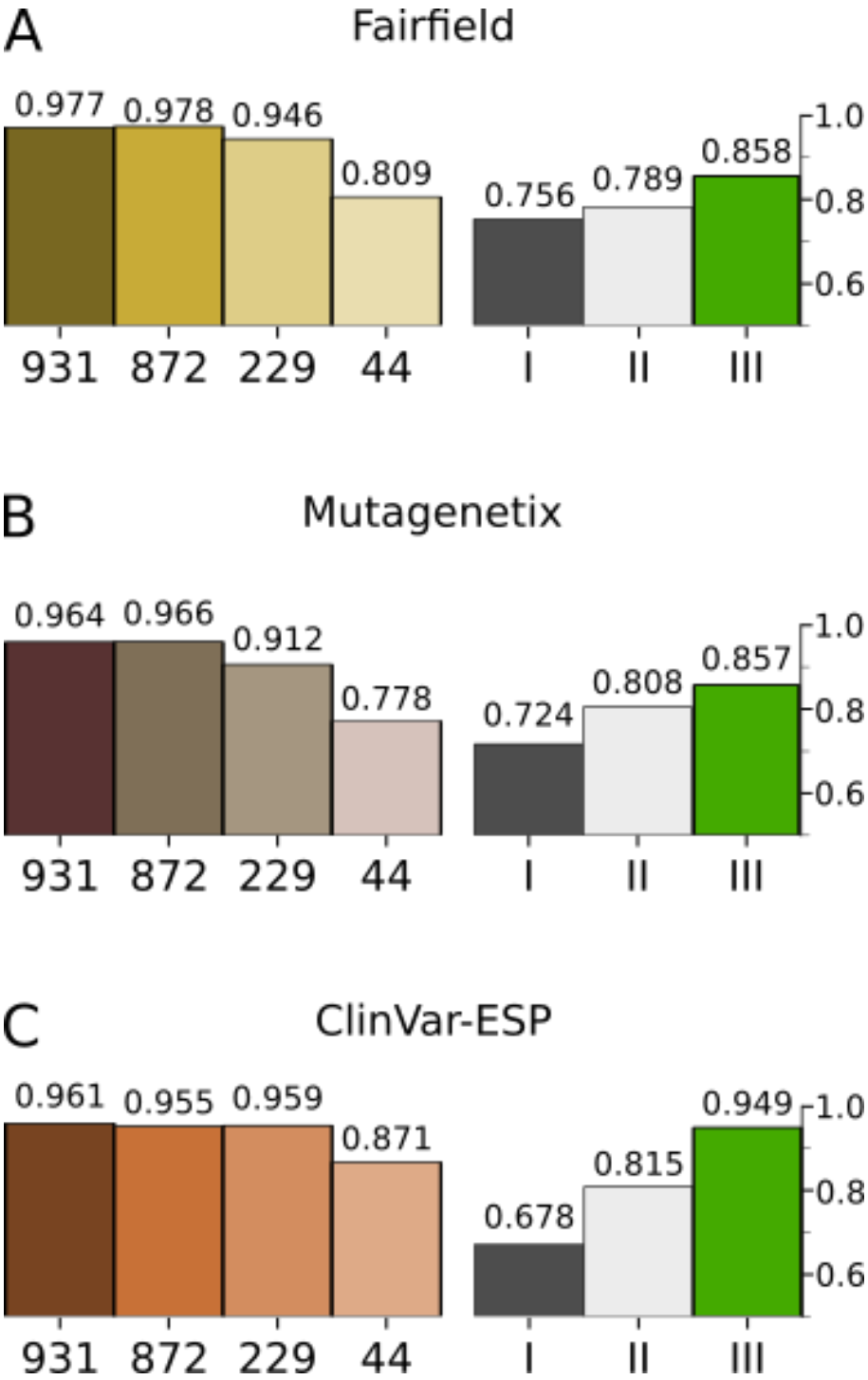


Figure 3: ROC-AUC scores of mCADD models evaluated on three different validation sets: the a) Fairfield, b) Mutagenetix and c) ClinVar-ESP data sets. The numbers below the bars indicate the number of annotations used during model training. Roman numbers indicate: I) the glire-PhastCons score, II) the vertebrate PhastCons score, and III) the hCADD* score. The numbers above the bars show the exact ROC-AUC of that particular model and validation set combination.

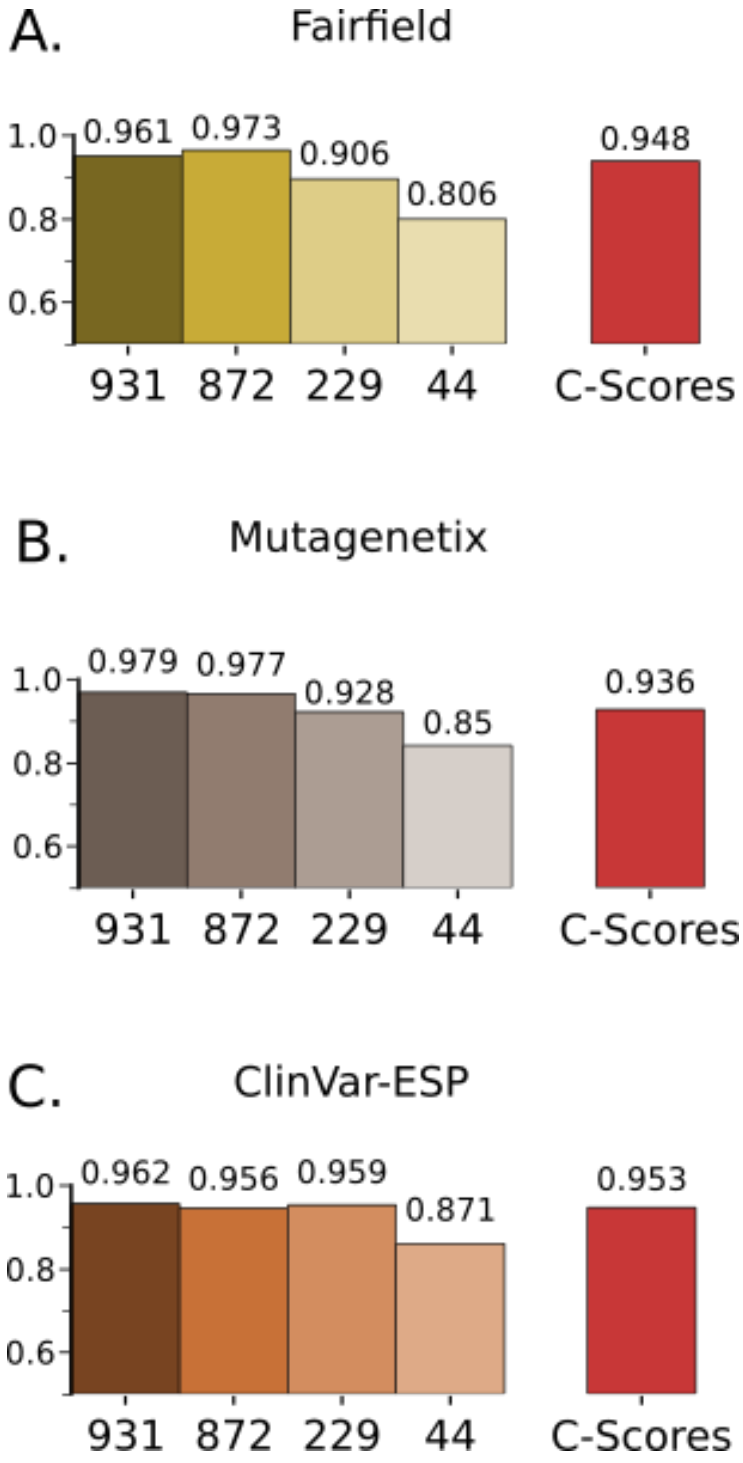


Figure 4: ROC-AUC scores of mCADD models and C-scores evaluated on three different validation sets (a) Fairfield, b) Mutagenetix, c) ClinVar-ESP) lifted from mouse to human. Arabic numbers underneath the bars indicate the number of annotations used for model training. The numbers above the bars show the exact ROC-AUC of that particular model and validation set combination.

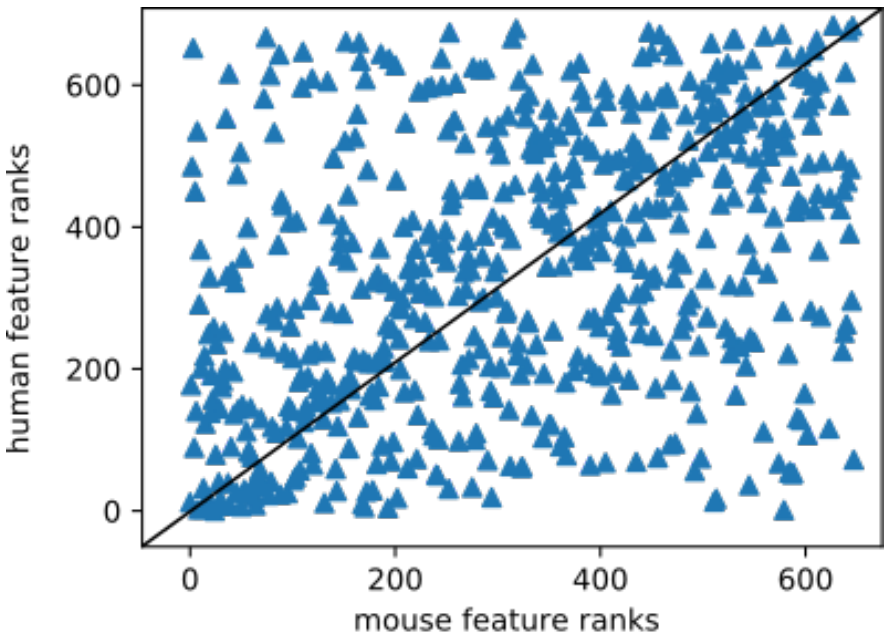


Figure 5: Comparing the ranks of the absolute weights assigned to annotations when training mCADD (horizontal axis) with those when training hCADD (vertical axis). A lower rank indicates an annotation with larger impact on the log-odds of a model.

2.3.4. Species-specific CADD model improves performance

To learn whether it is necessary to develop a mouse-specific model, we additionally lifted all three validation data sets from mm10 to GRCh37 and annotated the variants with CADD v.1.3 C-scores. We took care to only lift variants which have the same reference allele, thus displaying the same nucleotide substitution. Some variants could not be lifted due to a missing homozygous region. Negative samples were more often not lifted than positive ones, i.e. the Fairfield data set loses 50 negative samples and 27 positive ones, the Mutagenetix data set loses 235 positive and 398 negative samples, and for the ClinVar-ESP data set we had to omit 5 positive sample and 103 negative ones, due to the requirement of having the same reference allele.

For the Fairfield data set, the performance of all mCADD models dropped due to the removal of 77 samples (see Figure 4.A). The C-scores perform between mCADD(229) and mCADD(872). For the Mutagenetix data set, the mCADD models did not suffer from the removal of 633 SNVs, instead all computed ROC-AUCs increased (Figure 4.B). The C-scores perform again between mCADD(229) and mCADD(872). For the Clinvar-ESP data set, the mCADD model performances are hardly affected (see Figure 4.C). Applied on the ClinVar-ESP data set, mCADD(229) performs better than C-scores. Taken together, the species-specific mCADD model outperform lifting variants to human and using the hCADD model to score the variants, especially if considered that not every SNV can be easily lifted.

2.3.5. Annotation weights are moderately correlated between mCADD and hCADD

We examined whether different annotations are used by mCADD and hCADD. The absolutes of weights, assigned to each annotation by the logistic regressor, were ranked and the ranks of 595 annotations with a non-zero weight in both models were plotted against each other (see Figure 5), having a Spearman's rank correlation of 0.4.

Top-ranking mCADD annotations are enriched in combinations of DNA secondary structure predictions of DNAshape [23] (see Table S4). Furthermore, predictions of intronic and intergenic regions seem to be important, together with the neutral evolution score of GERP++ (GERPN) [18].

Top-ranking hCADD annotations are PhastCons and PhyloP conservation scores, all based on different phylogenies. Of these, the most influential annotations are PhastCons scores based on a primate alignment [5], [19]. The second most important group of annotations are predictions on intronic regions.

The combination of primate-based PhastCons scores in hCADD with predicted VEP consequences indicating intronic and intergenic regions is similar to the combination of the same VEP consequences and the neutral evolution score of GERP++ in mCADD. From this, we conclude that the primate-based PhastCons scores are replaced by GERPN in mCADD.

Vertebrate-based PhastCons scores are ranked high for both mCADD and hCADD. Top ranked annotations in hCADD which are ranked low in mCADD are enriched in mammalian-based PhastCons and mammalian-based PhyloP scores. Vice versa, feature combinations with DNA secondary structure predictions are exclusively used by mCADD.

2.4. Discussion

We demonstrated the possibility of creating a CADD-based model for the mouse genome, capable of predicting the deleteriousness of variants. We created a model trained on mouse data (mCADD) and evaluated it on a held-out test set and validation sets of phenotype altering SNVs. We compared the performance of our model to that of other metrics, such as conservation scores and the variant prioritization tool SIFT, as well as to C-scores for which we lifted the annotated variant locations to the human genome. We also compared performances on mouse test set variants to deleteriousness estimates of human test set variants, a.o. scored with a human CADD model that we trained ourselves (hCADD). As a final approach we trained a model on human data and evaluated it on mouse data (hCADD*).

Performances of mCADD and hCADD were very similar, with the mouse model performing better on the hold-out test sets. In addition, validation on three experimentally annotated data sets showed that the mCADD model is clearly capable of prioritizing deleteriousness of SNVs. Scoring lifted variants with hCADD performed reasonably well on these validation data sets, but less so than mCADD, whereas the generic hCADD* model had a consistent performance between mCADD(229) and mCADD(44). Together, this shows the importance of generating species-specific models when more annotations are available than only sequence specific ones, especially when lifting is not an option.

Evaluating the trained models on variants located in different genomic regions, we observed that mCADD and hCADD display the same trend, with increasing performance from non-coding to coding variants, and the best performance for missense mutations. Strikingly, mCADD, hCADD as well as other metrics all performed poorly on synonymous variants within coding regions.

We further assessed the annotation weightings in the human and mouse models. Despite a moderate correlation, both models rely on different annotations. This may explain the poorer performance of hCADD when evaluated on mouse data sets (i.e. hCADD*). Among the most important annotations are different conservation scores and/or combinations of these scores with VEP consequence annotations. It seems that hCADD relies relatively more on conservation scores than mCADD, while mCADD puts more emphasis on DNA structure predictions.

2.4.1. Performance depends on genomic region

Previous studies indicated that performance of the CADD classifier is not constant over the entire genome [21], [22]. We also observed changing performances between the investigated genomic regions. This may be due to intrinsic differences in the SNVs, but it might also be due to a difference in the number of annotations between non-coding and coding regions. When evaluating the distribution of putative deleterious and benign SNVs across genomic regions (Table S2), we find an imbalance in class labels of the held-out test set, but these do not explain the changes in performance. A striking difference in performance is found between the translated missense variants and translated synonymous variants. Annotations that help to differentiate between positive and negative missense mutations, such as SIFT, are not available for synonymous mutations. Hence, the main predictors for translated synonymous SNVs are the same as those for non-coding regions, namely different conservation scores, suggesting that the lack of meaningful annotations available for synonymous and other mutations limits the performance.

Note that CADD models are trained with putative benign and deleterious variants, as derived from the ancestor genome, and not with variants for which their effect is experimentally established. Although training variants are proxies, the trained CADD models perform extremely well on the experimentally validated SNVs as shown by the good performance on the validation sets. Apparently, the training variants are informative, and we, consequently, believe that the performances on the held-out test set can be interpreted at least qualitatively.

Together, this makes us believe that differences in observed performance between genomic regions are due to intrinsic properties of these regions such as the number of available annotations. This does, however, influence the applicability of any CADD-like model to prioritize disruptive SNVs truly genome wide.

2.4.2. Models based on limited numbers of annotations can be predictive

One of the objectives of this study was to investigate the predictive power of CADD-like models in the case of incomplete annotation sets when compared to the human case. For that purpose, we defined four different sub annotation sets: all annotations (mCADD(931), hCADD(1000)), all but epigenetic and species-specific annotations (m/hCADD(872)), annotations including VEP's (m/hCADD(229)), and annotations including only conservations scores (m/hCADD(44)).

The general trend is that mCADD models perform worse with fewer annotations, on the held-out test set as well as on the three validation sets. This is most pronounced for variants within coding regions. Differences in performance between mCADD(931) and mCADD(872) are negligible. For the Fairfield and Mutagenetix validation sets, mCADD(872) even performs better. The biggest drop in performance is observed between mCADD(872) and mCADD(229), even though the performance of mCADD(229) on all three validation sets is still above ROC-AUC > 0.91. These results indicate that a reliable model can be built, even if only very few annotations are known. Moreover, if only conservation scores and sequence features are available, it is still possible to outperform individual conservation scores.

hCADD shows a similar, but lower, trend, although the performance of hCADD(872) improves over that of hCADD(1000) using all subsets of the held-out test set. One of the main differences between mCADD and hCADD is that when generating training variants, mCADD uses an evolutionary older ancestor genome than hCADD. Thus, the time window over which mouse-derived variants have experienced purifying selection is longer than in the human case. Equally, substitution rates for the simulated SNVs are derived from evolutionary more distant ancestors, resulting in a larger proportion of deleterious SNVs in mouse than in human data. The impact of

the evolutionary observed differences is, however, poorly understood and warrants further investigation.

2.4.3. Limited interpretability of scores mapped between different species

An established method to evaluate different alleles in the genome of any species is to compare them with known orthologous regions in other species for which annotations are known. Although annotating lifted variants with human-based C-scores worked well, evaluating the same variants with a species-specific model gave better results. In addition, not every variant position in the validation sets could be annotated by C-scores as they have to be located in sequences that can be aligned to human. Further, similar variants in different species may differ in the phenotype they cause. This has to be considered for any comparative genomic analysis [24].

2.5. Conclusions

We have shown that the CADD approach for prioritizing variants can be applied to non-human species, and that it is important to train species-specific models. Interestingly, not all original annotations used by CADD are necessary to achieve good performance: only conservation scores and VEP consequences of variants (the set of 229 annotations we explored) may suffice to make meaningful predictions. These annotations are available for many species. Nevertheless, if possible, adding additional annotations for coding regions will help to improve the trained models. Altogether, our work has shown that species-specific CADD models can be successfully trained, opening new possibilities for prioritizing variants in other less well-studied species.

2.6. Methods

2.6.1. Overview of the CADD approach

We construct a CADD model for mouse, mCADD, as well as a CADD model based on human data, here denoted by hCADD. In contrast to the original CADD approach, mCADD and hCADD are trained specifically on single nucleotide variants. We also construct a model trained on human data and evaluated it on mouse variants, which will be further referred to as hCADD*. The purpose of this model is to learn about the performance to be expected if one wants to evaluate variants for which no model exists and that cannot be lifted between genomes. The SNVs and their annotations used for hCADD and hCADD* originate from the data set used for CADD v.1.3. Annotations that are specific for insertions or deletions were removed from the data set. Briefly, the original CADD model [5] is trained to classify variants as belonging to the class of simulated or derived variants. To train the CADD model, simulated and derived variants were generated based on the human-chimpanzee ancestral genome and mutation rates derived from a 6-taxa primate alignment [25].

Derived variants are variant sites with respect to the ancestral genome that are fixed in the human lineage, or nearly fixed with a derived allele frequency of above 95% in the 1000 Genomes Project. Due to the purifying selection they experienced, derived variants are assumed to be depleted in deleterious variants.

Next to observed derived variants, variants are simulated that do not occur in the human lineage. Hence, simulated variants did not experience purifying selection, therefore fitness reducing variants are not depleted in this group. All variants are annotated with a large number of genomic features, ranging from sequence features, conservation scores, variant effect predictor annotations to epigenetic measurements.

2.6.2. Derived and simulated variants in mouse

Due to a lack of sufficient sequencing data of large, freely reproducing mouse populations, we focused on identifying differences between an inferred mouse-rat ancestral genome and the most recent mouse reference assembly (mm10) [26]. The mouse-rat ancestral genome is based on the EPO 17-eutherian-mammal alignments [25], [27], [28] (Figure S2) provided by Ensembl release 83 [29]. In total we observed 33,622,843 sites with a derived allele in the mouse reference that were not adjacent to another variant site.

To generate an equal number of simulated variants we made use of the CADD variant simulator [5]. Based on the mm10 reference, it uses an empirical model of sequence evolution derived from the EPO 17-eutherian-mammal alignments, with CpG di-nucleotide specific rates and locally estimated mutation rates within windows of 100kb. Only SNVs with a known ancestral site were selected. In this way, we generated 33,615,003 SNVs. The final dataset contains an equal number of simulated variants, equally divided over 11 folds (10 for cross-validation and training, the remaining for testing), yielding a total of 67,229,998 SNVs. Table S2 gives an overview of these SNVs and their distribution over different genomic regions.

2.6.3. Genomic annotations

An overview of all annotations that we assembled for mouse can be found in Supplementary Data 2,3. Histone modifications, transcription factor binding sites, DNAase Seq peaks and RNAseq expression measurements were downloaded from ENCODE [16]. The mm10.60way vertebrate alignment was retrieved from the UCSC Genome Browser [30]. This multiple sequence alignment was used to calculate four different PhyloP and PhastCons scores based on differently sized subalignments, in particular an 8-taxa Glire alignment, a 21-taxa Euarchontoglire alignment, a 40-taxa Placental alignment and a 60-taxa Vertebrate alignment (Figure S1). PhyloP and PhastCons scores were computed without taking the mouse reference sequence into account. Furthermore, information about regulatory motifs, micro-RNA predictions (microRNA binding [31], microRNA targets [32]) and chromatin state predictions (ChromHMM [33]) were taken into account. GERP++ neutral evolution and rejected substitution scores, GERP Elements scores and GERP Elements p-values were taken from [18] and mapped from mm9 to mm10 via CrossMap [34]. All 5-mer combinations of the 4 nucleotides were generated and based on that the DNA secondary structure was predicted for each 5-mer [23]. Differences in the predicted scores for the reference 5-mer and alternative 5-mer at the investigated positions were used as annotation. Summaries of consequences predicted by the Ensembl Variant Effect Predictor (VEP v.87 [27]) were used in combination with other annotations to create additional composite annotations (Table S3, Supplementary Data 2, Supplementary Note). Additional annotations that rely on a gene build such as the SIFT protein score, reference and alternative amino acid, variant position within a transcript and coding region are also generated by VEP v.87.

Human annotations were downloaded from the original CADD publication v.1.3. [5] (download: 17-2-2016). Annotations which are by definition only available for InDels were removed.

2.6.4. Annotation subsets

From the annotations, four subsets were created of decreasing size and increasing likelihood of availability in non-human species (see Supplementary Data 2 for a complete overview). The first set consists of all available annotations, i.e. 1,000 for hCADD, 931 for mCADD and 902 for hCADD*. The annotations used to train hCADD* are those which can be meaningfully compared between mouse and human. The second subset has 872 annotations. It excludes all epigenetic

annotations and species-specific ones, leaving annotations available for both mouse and human. The third subset incorporates 229 annotations, including conservation scores, nucleotide sequence features and VEP consequence/annotation combinations. Annotations specific for coding regions were excluded, with the exception of coding region-specific VEP consequence values. The fourth subset of 44 annotations can be entirely generated from the sequence information itself. It includes conservation scores and nucleotide sequence annotations, such as the GC% within a 75bp window upstream and downstream of the variant position.

2.6.5. Training and evaluating the mCADD model

The CADD model is centered on a logistic regressor trained to differentiate between simulated and derived variants. This was done using the logistic regression module of Graphlab v2.0.1 [35], the same tool the CADD authors have used since CADD v1.1. Before training we standardized the human and mouse data by dividing each feature by its standard deviation. We did not center the features, in order to preserve sparsity. The mouse data set was split into 11 partitions of equal size (6,111,818 SNVs). The 11th partition was used as held-out test set. On the remaining 10 partitions we performed 10-fold cross validation to determine the number of training iterations for the logistic regressor and the L2 regularization parameter. The cross validation results are shown in Table S3. The final model was trained on the joined ten partitions with a maximum number of 100 iterations and a regularization parameter set to 0.1.

To obtain the human held-out test set, we selected 2,851,642 SNVs. Similar to the mouse case, this amounts to every 11th SNV from those available in the CADD v.1.3 data set. The hCADD and hCADD* models are trained with a maximum number of 10 iterations and an L2 regularization parameter of 1, to keep the settings as similar as possible to CADD v.1.3.

All model performances were evaluated with the area under the receiver operating characteristic (ROC-AUC). Trained classifiers were assessed based on their performances on their respective held-out test sets. These sets were further divided according to the genomic regions from which each variant originates. An overview and description of the resulting 7 subsets can be found in Table S2.

We further evaluated the classifiers on three additional data sets: (i) 60 SNVs associated with changes in phenotype as obtained from an exome sequencing study of 91 mouse strains with Mendelian disorders (Fairfield data set) [36]; (ii) 481 N-ethyl-N-nitrosourea (ENU) induced SNVs (Mutagenetix data set) [37]; (iii) 9,348 variant sites lifted from the ClinVar-ESP validation set utilized in CADD v.1.3 (ClinVar-ESP data set) [5]. Similar to the training data, all data sets were standardized but not centered, using the scaling factors for each annotation which were obtained from the whole mouse data set.

Data for the Fairfield validation set is provided by Table S4 [38] of the Fairfield et al. publication. The Mutagenetix data set was provided by several labs and downloaded from the Mutagenetix data base [37], [39]. All data were checked for the reported reference allele and, in the case of uncertainty, manually verified with the records on the website. If the reported allele could not be found in close proximity of the reported genomic location, the variant was discarded. Both the Fairfield and Mutagenetix validation sets contain phenotype altering SNVs, therefore all of these were considered as potentially deleterious without differentiating between the exact nature of the phenotype change (positive data set). To find an equal number of variants that can be used as a negative data set, we made use of SNVs identified in 36 mouse strains from the Wellcome Trust Sanger's Mouse Genomes Project [15], filtered for an allele frequency (AF) ≥ 90%. We sampled to have a matching number of negative SNVs for both data sets, we took care that the proportions of

transcribed, synonymous and non-synonymous mutations are the same among the positive and negative SNVs.

The ClinVar-ESP data set contains curated variants from the ClinVar database [14] that were identified to have a pathogenic effect in human. As a negative set (5,635 SNVs), variants from the Exome Sequencing Project (ESP) [40] were selected with a derived allele frequency of ≥ 5%. We lifted the variants from GRCh37 to mm10 and selected SNVs which introduce the same amino acid substitution or stop codon change in human and mouse.

2.6.6. Analysis of model weights

The logistic regressor assigns weights (betas) to each annotation used for training. These weights indicate the effect of one unit change on the log odds of success of the trained model. A zero weight implies that the annotation is not used. We compared the weights assigned to each annotation by mCADD and hCADD to derive information about annotations of general importance for CADD-like models. As different regularization terms were applied in hCADD and mCADD, causing the beta's to be on different scale, we compared ranks instead of weights. Ranks were computed for non-zero beta's and based on the absolute weight. Annotations of mCADD and hCADD were compared with each other when they have a non-zero weight in both models. Three types of annotations were not identical between mouse and human, but considered comparable:

- Primate-based PhastCons&PhyloP [19], [20] scores in hCADD were compared with glire-based PhastCons&PhyloP scores of mCADD. These are the smallest alignments used to compute conservation scores in both species.
- Mammalia based PhastCons&PhyloP scores in hCADD were compared to scores based on a placentalia alignment for mCADD.
- CHROMHMM [33] chromatin state predictions were mapped based on the overlap of their predicted consequences in human and mouse.

2.7. Declarations / Statements

2.7.1. Availability of data and materials

Data and scripts to reproduce the results can be downloaded from the following link. http://www.bioinformatics.nl/mCADD/mCADD_Data_and_Scripts.tar.gz

2.7.2. Funding

This research was funded by the STW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality. This study was financially supported by NWO-TTW and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs-Norsvin. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.7.3. Acknowledgements

We thank Martin Kircher for providing data sets and answering various questions regarding CADD. We also thank our collaborators Martijn Derks, Mirte Bosse, Hendrik-Jan Megens and Martien Groenen for valuable discussions. Last but not least, we thank the anonymous reviewers for critical suggestions which helped to improve the manuscript.

Bibliography

[1] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Res.*, vol. 11, no. 5, pp. 863–874, 2001.

[2] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[3] M. Hecht, Y. Bromberg, and B. Rost, "Better prediction of functional effects for sequence variants," *BMC Genomics*, vol. 16, no. 8, p. S1, 2015.

[4] Y. Choi and A. P. Chan, "PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics*, vol. 31, no. 16, pp. 2745–2747, 2015.

[5] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants.," *Nat. Genet.*, vol. 46, no. 3, pp. 310–5, 2014.

[6] G. R. S. Ritchie, I. Dunham, E. Zeggini, and P. Flicek, "Functional annotation of noncoding sequence variants," *Nat. Methods*, vol. 11, no. 3, pp. 294–296, 2014.

[7] H. A. Shihab *et al.*, "An integrative approach to predicting the functional effects of non-coding and coding sequence variation," *Bioinformatics*, vol. 31, no. 10, pp. 1536–1543, 2015.

[8] B. Guiko, M. J. Hubisz, I. Gronau, and A. Siepel, "Probabilities of fitness consequences for point mutations across the human genome," *Nat. Genet.*, vol. 47, no. 3, pp. 276–283, 2015.

[9] Y. F. Huang, B. Gulko, and A. Siepel, "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data," *Nat. Genet.*, vol. 49, no. 4, pp. 618–624, 2017.

[10] S. Balasubramanian *et al.*, "Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes," *Nat. Commun.*, vol. 8, no. 1, 2017.

[11] H. E. Abboud *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2017.

[12] T. 1000 G. P. Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, 2015.

[13] M. Peplow, "The 100 000 genomes project," *BMJ*, vol. 353, pp. 2018–2020, 2016.

[14] M. J. Landrum *et al.*, "ClinVar : public archive of interpretations of clinically relevant variants," vol. 44, pp. 862–868, 2016.

[15] T. M. Keane *et al.*, "Mouse genomic variation and its effect on phenotypes and gene regulation," *Nature*, vol. 477, no. 7364, pp. 289–294, 2011.

[16] I. Dunham *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[17] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites," *Genome Biol.*, vol. 11, no. 8, 2010.

[18] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Comput. Biol.*, vol. 6, no. 12, 2010.

[19] A. Siepel *et al.*, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.," *Genome Res.*, vol. 15, no. 8, pp. 1034–50, 2005.

[20] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, 2010.

[21] Y. Itan *et al.*, "The mutation significance cutoff: gene-level thresholds for variant predictions," *Nat. Methods*, vol. 13, no. 2, pp. 109–110, 2016.

[22] C. A. Mather *et al.*, "CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel," *Genet. Med.*, vol. 18, no. 12, pp. 1269–1275, 2016.

[23] T. Zhou *et al.*, "DNAshape : a method for the high-throughput prediction of DNA structural features on a genomic scale," vol. 41, pp. 56–62, 2013.

[24] B.-Y. Liao and J. Zhang, "Null mutations in human and mouse orthologs frequently result in different phenotypes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 19, pp. 6987–6992, 2008.

[25] B. Paten, J. Herrero, K. Beal, S. Fitzgerald, and E. Birney, "Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs," *Genome Res.*, vol. 18, no. 11, pp. 1814–1828, 2008.

[26] D. M. Church *et al.*, "Lineage-specific biology revealed by a finished genome assembly of the mouse," *PLoS Biol.*, vol. 7, no. 5, 2009.

[27] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor.," *Genome Biol.*, vol. 17, no. 1, p. 122, 2016.

[28] B. Paten *et al.*, "Genome-wide nucleotide-level mammalian ancestor reconstruction," *Genome Res.*, vol. 18, no. 11, pp. 1829–1843, 2008.

[29] D. R. Zerbino *et al.*, "Ensembl 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, 2018.

[30] M. L. Speir *et al.*, "The UCSC Genome Browser database: 2016 update," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D717–D725, 2016.

[31] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander, "The microRNA.org resource: targets and expression," *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, pp. 149–153, 2008.

[32] B. P. Lewis, I. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, pp. 787–798, 2003.

[33] J. Ernst and M. Kellis, "ChromHMM: Automating chromatin-state discovery and characterization," *Nat. Methods*, vol. 9, no. 3, pp. 215–216, 2012.

[34] H. Zhao, Z. Sun, J. Wang, H. Huang, J. Kocher, and L. Wang, "CrossMap : a versatile tool for coordinate conversion between genome assemblies," vol. 30, no. 7, pp. 1006–1007, 2014.
[35] Turi, "Graphlab create." [Online]. Available: https://turi.com/index.html. [Accessed: 14-Mar-2017].
[36] M. Kircher *et al.*, "Exome sequencing reveals pathogenic mutations in 91 strains of mice with Mendelian disorders," *Genome Res.*, vol. 25, no. 7, pp. 948–957, 2015.
[37] T. Wang *et al.*, "Real-time resolution of point mutations that cause phenovariance in mice," *Proc. Natl. Acad. Sci. U. S. A.*, pp. E440–E449, 2015.
[38] "Fairfield et al. Supplementary Table 4." [Online]. Available: https://genome.cshlp.org/content/suppl/2015/04/23/gr.186882.114.DC1/Supplemental_Table_4.xlsx. [Accessed: 22-May-2018].
[39] "Mutagenetix Phenotypic Mutations." [Online]. Available: https://mutagenetix.utsouthwestern.edu/phenotypic/phenotypic_list.cfm. [Accessed: 26-Mar-2018].
[40] W. Fu *et al.*, "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants," *Nature*, vol. 493, no. 7431, pp. 216–220, 2013.

2.8. Appendix – Supplementary Data

Supplementary files "*supplementary_data2-4.xlsx"* are available online under:

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2337-5

2.8.1. Supplementary Note

2.8.1.1. *Annotation pre-processing*

To train mCADD and hCADD models only SNV were considered. Differences between the inferred ancestor genome and the mouse reference were utilized as negative class for training. Differences were considered when they were not adjacent to another site that was different between the ancestor and reference. These mutations are directed back in time while simulated variants are orientated forward in time. Therefore annotations that are sensitive to these differences have to be swapped in the set of derived variants. Namely, the nucleotide reference and alternative columns (*Ref, Alt*), the amino acid substitutions (*nAA, oAA*) and the variant effect consequence predictions made by the ENSEMBL Variant Effect Predictor v87 for the labels (*STOP_Gained, STOP_LOST*).

motifEHIPos, GerpRS, SIFTval, GerpRSpval, mirSVR-Score, mirSVR-E, mirSVR-Aln, targetScan, Expression, DNAseSig, H3K27ac, H3K4me1, H3K4me3, tOverlapMotifs, motifDist, motifECount, motifEScoreChng, TFBS, TFBS-Peak, TFBSPeaksMax, cDNApos, relcDNApos, CDSpos, relCDSpos, prot-Pos, relprotPos, Dst2Splice, Grantham

The following annotations were mean imputed based on the mean of the simulated variants:

GC, CpG, dnaRoll, dnaProT, dnaMGW, dnaHelT, GerpN, GerpS, GerpRS, euaPhCons, euaPhyloP, gPhCons, gPhyloP, minDistTSS, minDistTSE, plaPhCons, plaPhyloP, verPhCons, verPhyloP

For the following annotations, another category (UD = undened) was introduced to indicate missing values:

Domain, Dst2SplType, SIFTcat, oAA, nAA

Missing values in the annotation (*isTv*) were replaced by 0.5.

For the set of following annotations, an indicator feature was created which is set to 0 if the annotation is dened and set to 1 if undefined:

Dst2SplType_ACCEPTOR, Dst2SplType_DONOR, mirSVR-Score, targetScan, cDNApos, CDSpos, protPos, SIFTval, Grantham

The annotations (*minDistTSE ,minDistTSS*) were capped at 10000.

The following annotations were log-transformed:

minDistTSE ,minDistTSS, GerpRS

All categorical annotations were OneHotEncoded. Further annotation combinations were created. Namely, all possible combinations of *Ref* and *Alt*, representing an annotation for each possible nucleotide substitution. The same was done for *nAA* and *oAA*, thus there is one annotation for each possible amino acid substitution. Lastly, combinations of the set of the following annotations were made with each of the 15 summarized consequences (Supplementary Data 2) of the Ensembl Variant Effect Predictor.

cDNApos, CDSpos, Dst2Splice, GerpS, GerpN, plaPhCons, plaPhyloP, minDistTSE, minDistTSS, euaPhCons, euaPhyloP, protPos, relcDNApos, relCDSpos, rel-protPos, verPhCons, verPhyloP, dnaHelT, dnaMGW, dnaProT, dnaRoll, gPhCons, gPhyloP

2.8.2. Supplementary Tables

Table S1: VEP consequences are summarized in 15 categories. If multiple annotations exist for the same variant, the consequence is selected according to the displayed hierarchy, with STOP-GAINED being the most important and UNKNOWN the least important category.

Hierarchy	Abbreviation	VEP Consequence categories
1	SG	STOP-GAINED
2	CS	CANONICAL-SPLICE
3	NS	NON-SYNONYMOUS
4	SN	SYNONYMOUS
5	SL	STOP-LOST
6	S	SPLICE-SITE
7	U5	5PRIME-UTR
8	U3	3PRIME-UTR
9	R	REGULATORY
10	IG	INTERGENIC
11	NC	NONCODING-CHANGE
12	I	INTRONIC
13	UP	UPSTREAM
14	DN	DOWNSTREAM
15	O	UNKNOWN

Table S2: This table gives a description about the genomic regions which were selected to evaluate the mCADD and hCADD models. Underneath the Genomic region, the total number of SNVs located in that region is displayed. H=Human, M=Mouse.

Genomic Region Total number SNV	Description	Class distribution Human	Class distribution Mouse
entire genome H:31,368,062, M:67,229,998	randomly selected SNVs taken from the entire genome.	Derived: 0.5 Simulated: 0.5	Derived: 0.5 Simulated: 0.5
not transcribed H:30,592,093, M:64,278,844	randomly selected SNVs which are located outside of known transcript regions.	Derived: 0.5 Simulated: 0.5	Derived: 0.5 Simulated: 0.5
Transcribed H:775,969, M:2,951,154	randomly selected SNVs which are located in known transcript regions.	Derived: 0.4 Simulated: 0.6	Derived: 0.46 Simulated: 0.54
transcribed not translated H:461,057, M:1,684,821	randomly selected SNVs which are located in transcript regions but not translated. (5'UTR, 3'UTR, Intron)	Derived: 0.47 Simulated: 0.53	Derived: 0.5 Simulated: 0.5
translated H:314,912, M:1,266,333	randomly selected SNVs which are located in known translated regions (Exon).	Derived: 0.29 Simulated: 0.71	Derived: 0.42 Simulated: 0.58
translated synonymous H:126,103, M:625,183	randomly selected SNVs which are located in translated regions but do not code for a missense annotations with an associated SIFT value.	Derived: 0.41 Simulated: 0.59	Derived: 0.62 Simulated: 0.38
Translated missense H:188,809, M:641,150	randomly selected SNVs in translated regions that have a missense annotation with an associated SIFT value.	Derived: 0.21 Simulated: 0.79	Derived: 0.23 Simulated: 0.77

Table S3: 10-fold cross validation performance of mCADD models. Each row is showing a different number of iterations, each column a different L2-penalization.

Iteration | L2-Penalization	**0.1**	**1**	**10**
10	Mean: 0.623 Std: 0.01	Mean: 0.625 Std: 0.011	Mean: 0.626 Std: 0.009
100	Mean: 0.668 Std: 0.001	Mean: 0.634 Std: 0.104	Mean:0.667 Std: 0.003
1000	Mean: 0.638 Std: 0.06	Mean: 0.638 Std: 0.076	Mean: 0.653 Std: 0.042

Table S4: Top performing predictors in hCADD and mCADD

top 10 mCADD	**top 10 hCADD**	**top 10 hCADD and mCADD**	**hCADD>500 and 100>mCADD**	**hCADD<100 and 500<mCADD**
GerpN IGxGerpN dnaRoll SIFTval IxGerpN IGxdnaRoll dnaMGW IxdnaRoll verPhCons GC	priPhCons mamPhCons verPhCons verPhyloP mamPhyloP priPhyloP IxpriPhCons IGxpriPhCons GerpS IxverPhyloP	verPhCons	UPxdnaMGW DNxdnaMGW RxdnaHelT IxdnaRoll	IGxmamPhyloP IGxmamPhCons RxmamPhCons oAAxUD IND_protpos mamPhCons nAAxUD IND_CDSpos

2.8.3. Supplementary Figures

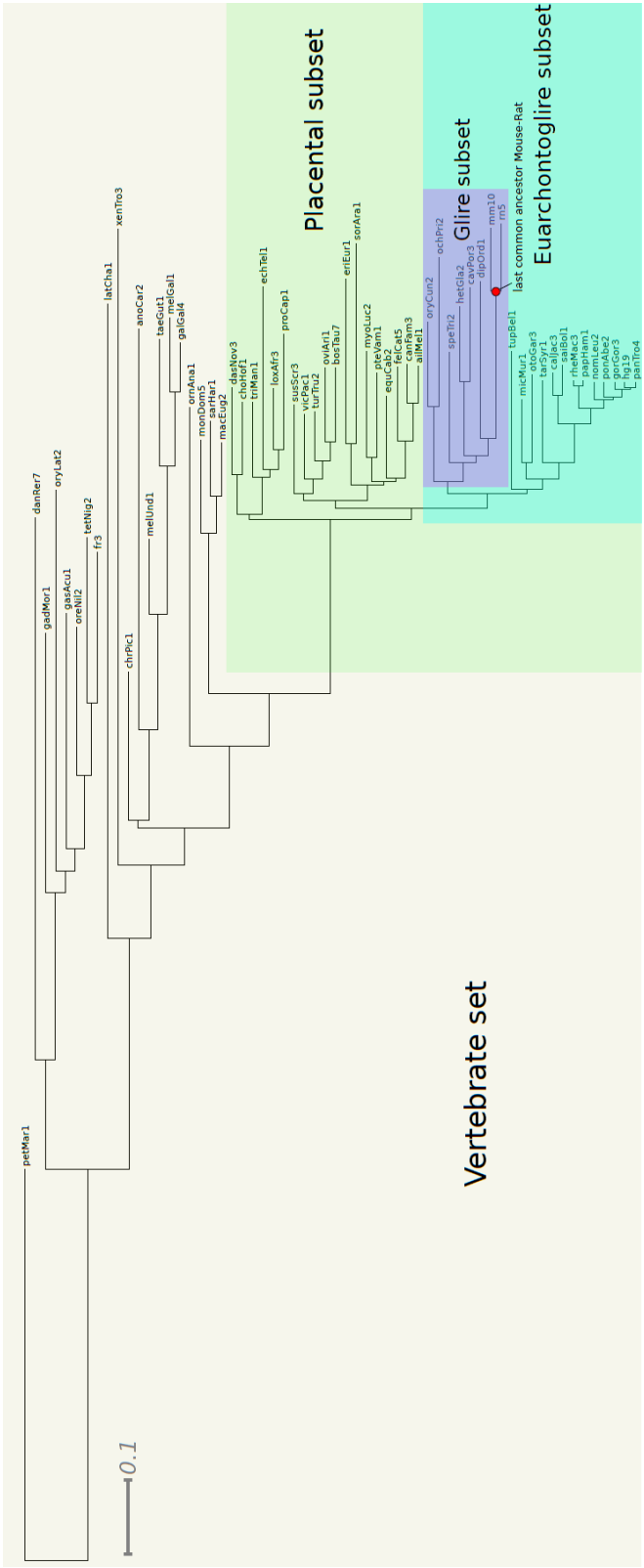


Figure S1: Phylogenetic tree, displaying the Vertebrate, Placental, Euarchontoglire and Glire sets which were used to compute PhastCon and PhyloP conservation scores. Furthermore, the last common ancestor between Mouse and Rat is indicated.

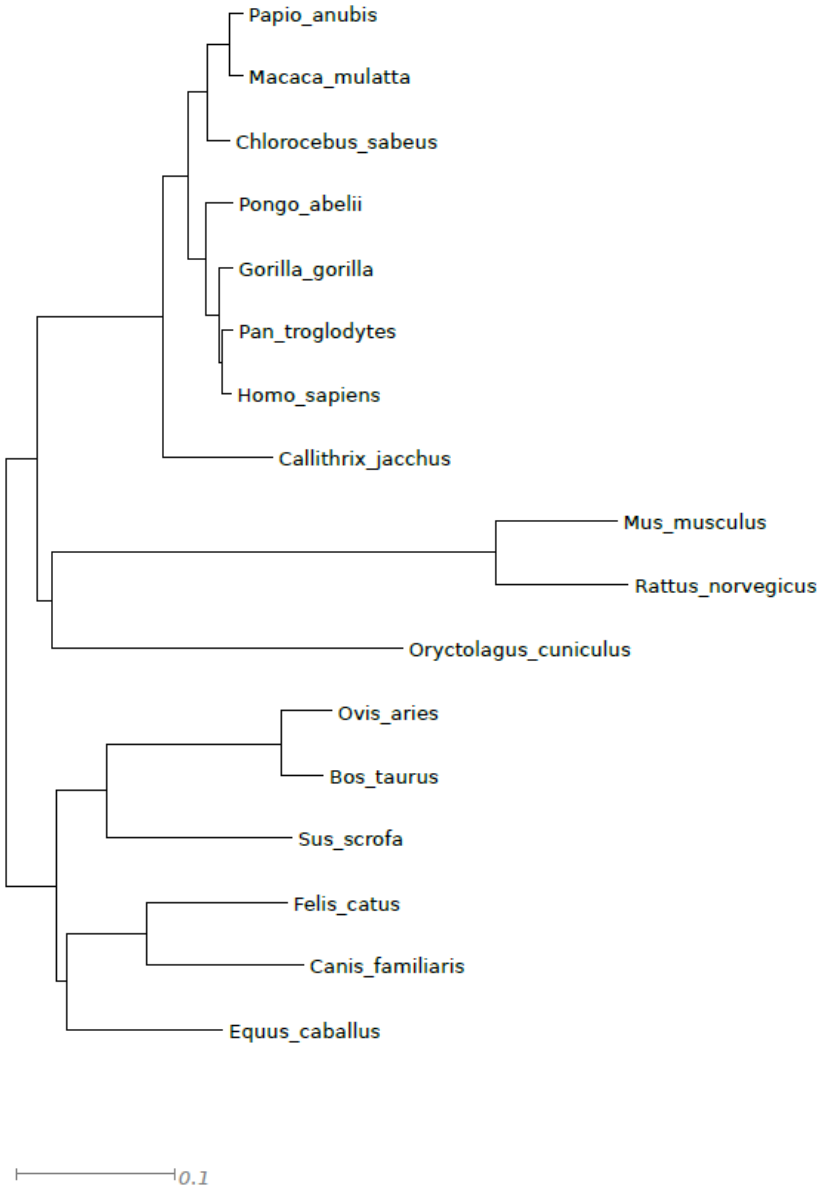


Figure S2: Phylogenetic tree, displaying the taxa used in the 17-eutherian mammal EPO alignment. That alignment was used to infer the mouse ancestral sequence and to derive substitution rates to simulate variants.

3. pCADD: SNV prioritisation in *Sus scrofa*

Christian Groß

Martijn Derks

Hendrik-Jan Megens

Mirte Bosse

Martien A.M. Groenen

Marcel Reinders

Dick de Ridder

This chapter is published as: C. Groß, M. Derks, HJ Megens, et al. "pCADD: SNV prioritisation in *Sus scrofa*." *Genet Sel Evol* 52, 4 (2020). https://doi.org/10.1186/s12711-020-0528-9

3.1. Abstract

Background: In animal breeding, identification of causative genetic variants is of major importance and high economical value. Usually, the number of candidate variants exceeds the number of variants that can be validated. One way of prioritizing probable candidates is by evaluating their potential to have a deleterious effect, e.g. by predicting their consequence. Due to experimental difficulties to evaluate variants that do not cause an amino-acid substitution, other prioritization methods are needed. For human genomes, the prediction of deleterious genomic variants has taken a step forward with the introduction of the combined annotation dependent depletion (CADD) method. In theory, this approach can be applied to any species. Here, we present pCADD (p for pig), a model to score single nucleotide variants (SNVs) in pig genomes.

Results: To evaluate whether pCADD captures sites with biological meaning, we used transcripts from miRNAs and introns, sequences from genes that are specific for a particular tissue, and the different sites of codons, to test how well pCADD scores differentiate between functional and non-functional elements. Furthermore, we conducted an assessment of examples of non-coding and coding SNVs, which are causal for changes in phenotypes. Our results show that pCADD scores discriminate between functional and non-functional sequences and prioritize functional SNVs, and that pCADD is able to score the different positions in a codon relative to their redundancy. Taken together, these results indicate that based on pCADD scores, regions with biological relevance can be identified and distinguished according to their rate of adaptation.

Conclusions: We present the ability of pCADD to prioritize SNVs in the pig genome with respect to their putative deleteriousness, in accordance to the biological significance of the region in which they are located. We created scores for all possible SNVs, coding and non-coding, for all autosomes and the X chromosome of the pig reference sequence Sscrofa11.1, proposing a toolbox to prioritize variants and evaluate sequences to highlight new sites of interest to explain biological functions that are relevant to animal breeding.

3.2. Background

Since humans started breeding animals, a key challenge has been to control the inheritance of traits. In farm animals, genetic gain has been achieved using pedigree information and statistical models. Since the introduction of genomic selection (GS) [1], breeding is transitioning from selecting animals based on visual inspection and pedigree data to approaches that exploit genetic information. However, given the complexity of genomes and the generally low level of knowledge about the relation between genotype and phenotype, undesirable alleles may accumulate, through genetic hitchhiking or genetic drift [2], [3] because of the small effective population size in livestock breeds under artificial selection.

Recent approaches incorporate whole-genome sequence data to improve genetic predictions. Because the number of tested single nucleotide variants (SNVs) is larger in whole-genome sequence data compared to array-based assays, truly causal genetic variants are more likely to be identified. While the use of whole-genome sequence data has improved genetic prediction, the improvements fall short of expectation and yield only moderate performance increases [4], [5], partly due to the inclusion of noise. Therefore, current strategies involve pre-weighting of potential candidate SNVs that have a higher probability of being causal. Several methods have been developed to score variants according to their putative deleteriousness and identify those that may have a detrimental effect on the fitness of individuals. Well-known variant prioritization tools include SIFT [6], PolyPhen2 [7], SNAP2 [8] and Provean [9]. However, these are limited to scoring

(non-synonymous) variants in coding regions. In contrast, the combined annotation dependent depletion (CADD) [10] model that was developed to investigate SNVs in human populations, can score variants at any location in the genome. CADD is comparable to methods such as fitCons [11] and Linsight [12]: it captures signals of evolutionary selection across many generations and combines this with annotations—genomic features, epigenetic data, other predictors etc.—to estimate a deleteriousness score for a given variant. While CADD and similar models are well established and used to predict the effects of variants in the human genome [13]–[18], to date, they have not been applied to non-human species. In recent work [19], we applied CADD to mouse, and studied the effect of having a limited number of annotations, which is expected for non-model species, compared to the human case. The results demonstrated that applying the CADD methodology to non-human species is valid and powerful.

Here, we introduce pCADD (p for pig), a model based on the CADD methodology to create scores for the prioritisation of SNVs with respect to their putative deleteriousness in the genomes of wild and domesticated pigs (Sus scrofa). The aim of this paper is to assess the ability of pCADD to prioritize individual SNVs and genomic regions relative to their biological function. The ability of pCADD to score any SNV in the entire pig genome with respect to its predicted deleteriousness helps researchers and breeders to evaluate (newly) observed SNVs and rank potentially harmful SNVs that are propagated by breeding.

3.3. Methods

Briefly, the CADD model, which is a logistic regressor, assigns a deleteriousness score to a SNV based on a set of 867 genomic annotations such as DNA secondary structure, conservation scores, protein function scores and many more (see Additional file 1 and Additional file 2: Table S1). Model parameters are fitted based on a large training set, containing two classes of SNVs: derived (proxy benign/neutral) and simulated (proxy deleterious) SNVs. The set of derived SNVs is generated by identifying (nearly) fixed alleles in the species of interest that differ from those of a reconstructed ancestral genome (Figure 1a). Proxy deleterious SNVs are simulated de novo mutations, which have not experienced any selection, thus deleterious variants are not depleted in this set (Figure 1b, c).

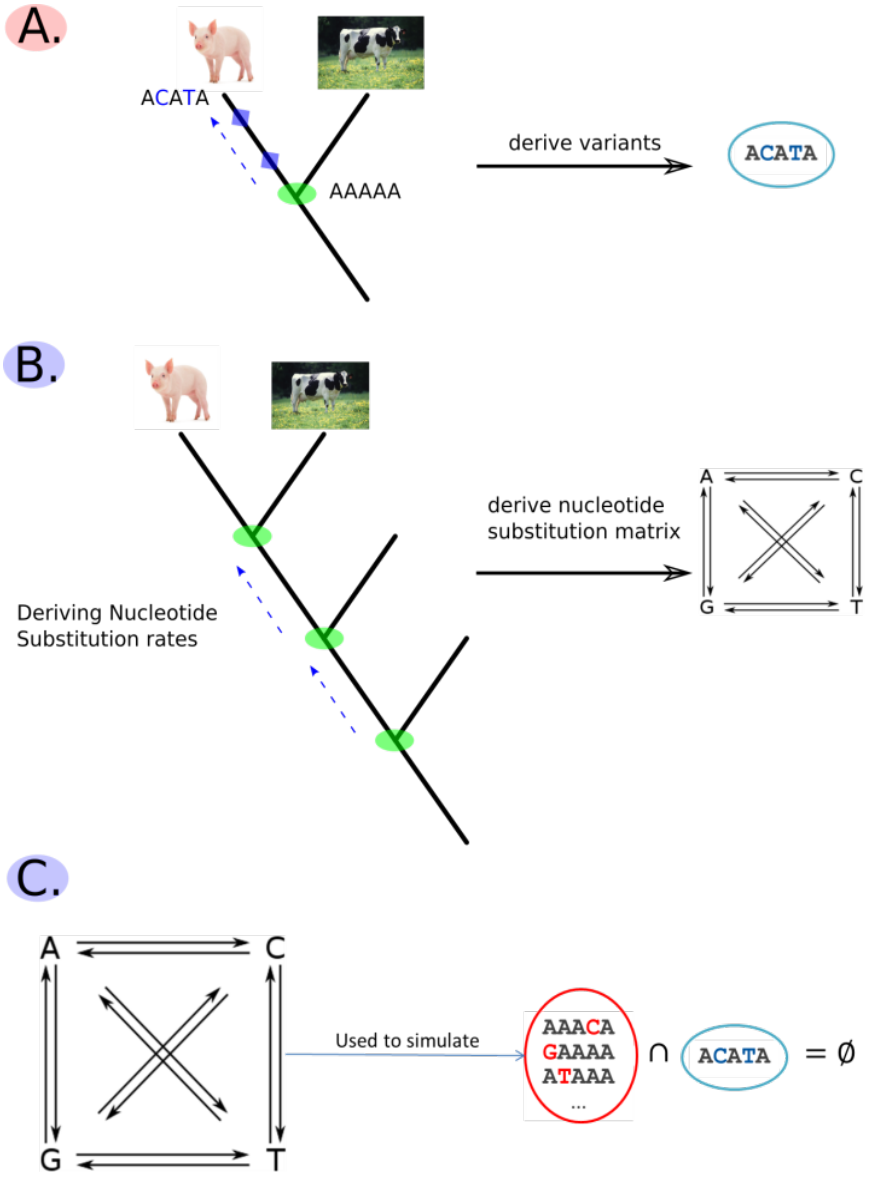


Figure 1: A. Fixed or almost fixed differences between an inferred ancestor sequence and the investigated pig population are used as proxy benign/neutral SNVs. B. Simulation, first step: differences between differently deep ancestor sequences are identified and substitution rates are derived. C. Simulation, second step: the derived substitution rates are used to simulate de novo variants that have not experienced any selection and therefore are not depleted in the number of deleterious variants.

With the pCADD model, every position in the pig genome can be scored with respect to its predicted deleteriousness. To differentiate more easily those SNVs that are potentially of interest, we created a PHRED-like score, which is similar to that in the original CADD approach [10]. To this end, the outcomes of the logistic regressor for all variants are ordered and transformed. The pCADD score is a log-rank score that ranges from ~95 to 0, with higher scores indicating more deleterious variants. The top 1% and 0.1% highest scored SNVs have a pCADD score higher than 20 and 30, respectively, thus the most deleterious variants are differentiated from the likely neutral ones. In the following, we describe the data used to train the pCADD model and demonstrate its use by performing several analyses.

3.3.1. Training and test set construction

To create the set of derived variants, which consists of putatively benign/neutral variants, we identified (nearly) fixed alleles in a pig population that differ from those of the reconstructed ancestral genome of pig, cow and sheep (Figure 1a, Sus scrofa [20], Bos taurus [21] , Ovis aries [22]]). These alleles have become fixed in the pig population due to genetic drift or positive selection, thus they are depleted in deleterious variants and can be assumed to have a benign or neutral effect. The ancestral sequence was obtained from the 25-eutherian-mammals EPO (Enredo, Pecan, Ortheus) [23], [24] multiple alignment files (MAF), downloaded from the Ensembl v.91 database. To avoid errors due to misaligned InDels, only SNVs that are not adjacent to another variant site, between the pig population and the inferred ancestor, were retained. The pig population used in our study included 384 individuals, representing 36 breeds, e.g. Asian and European, wild, commercial and local breeds (see Additional file 2: Table S2). For each site in the inferred ancestor, we selected an allele when its frequency was higher than 0.9 in the pig population and when it differed from the ancestral allele. Because the population includes pigs from many breeds, the number of functional variants that may have reached fixation due to founder effects in individual populations is limited. In addition, we removed sites that carry an allele at a frequency higher than 0.05 in the population and for which the alternate allele is equal to the ancestral allele. To simulate variants for the proxy deleterious set, substitution rates were derived from observed differences between more distant ancestors of pig (Figure 1b, c). In particular, rates for nucleotide substitutions and CpG sites in window sizes of 100 kb were computed based on the inferred substitutions between the ancestral sequences of pig-cow, pig-horse and pig-dog. Only SNVs that were located at a site with a known ancestral allele of the pig-cow-sheep ancestor were simulated. These SNVs are de novo mutations that have a larger than uniform chance, with respect to other de novo mutations, to occur in the populations. Although these variations may have never occurred by chance along the evolutionary branch of pig, they may have also been actively selected against. In other words, these random mutations have a greater chance of being deleterious than benign [25], therefore the set of simulated variants is expected to be enriched in deleterious variants in comparison to the derived proxy benign/neutral set.

In total, 61,587,075 proxy benign/neutral SNVs were derived and a similar number of SNVs was simulated. To form the training and test sets, the dataset was randomly split into two sets with an equal number of samples from both classes. The training dataset contained 111,976,500 SNVs whereas the test set consisted of 11,197,650 SNVs. To assess the dependency on the genomic location of the variants, the test set was split into six overlapping subsets: (i) intergenic (non-cDNA) variants; (ii) all transcribed sites (cDNA); (iii) transcribed but not translated sites (5'UTR5, 3'UTR3 and introns); (iv) coding regions; (v) synonymous SNVs in coding regions and (vi) non-synonymous SNVs in coding regions.

3.3.2. Variant annotation

Genomic annotations were obtained from the Ensembl Variant Effect Predictor (VEP v91.3) database [26] and supplemented by PhyloP [27], PhastCons [28] and GERP [29] conservation scores as well as Grantham [30] amino-acid substitution scores and predictions of secondary DNA structure (DNAshape) [31].

VEP-predicted consequences of SNVs were summarised in 14 categories. They were either used directly or combined with other data to create composite annotations (see Additional file 1 and Additional file 2: Table S3). Annotations that rely on a gene build, such as the SIFT protein score, reference and alternative amino-acid, variant position within a transcript and coding region were also used.

PhyloP and PhastCons scores are based on three differently sized multiple species alignments: a 6-taxa laurasiatheria, a 25-taxa eutherian-mammals and a 100-taxa vertebrate alignment. The laurasiatheria and eutherian-mammals alignments were downloaded from Ensembl [32] v91 whereas the 100-taxa vertebrate alignment was downloaded from UCSC [33], [34] (December 29, 2017). Next, PhyloFit [35] phylogenetic models were created for the laurasiatheria and eutherian-mammals alignments to compute PhastCons and PhyloP scores for pig. PhyloFit models for the 100-taxa vertebrate alignment were downloaded from the UCSC genome browser and used to compute PhastCons and PhyloP scores. PhastCons and PhyloP scores based on the 6- and 25-taxa alignments were directly computed for pig, while the scores for the 100-taxa alignment had to be first computed for the human reference GRCh38 and then mapped to Sscrofa11.1 using CrossMap [36]. To avoid a positive bias in predictive power in favour of PhastCons and PhyloP scores, the pig sequence was excluded from the generation of both sets of scores. Genomic evolutionary rate profiling (GERP) neutral evolution, GERP conservation, GERP constrained element and GERP constrained element p-values were retrieved from Ensembl91 using a custom Perl script.

Predicted differences in the secondary DNA structure between reference and alternative alleles were added as annotations to the dataset, as computed by DNAshape [31]: minor gap width (MGW), Roll, propeller twist (ProT) and helix twist (HelT).

After computing all annotation combinations, imputing missing values and recoding all categorical values to binary variables (see Additional file 1), the final number of features was equal to 867. Each feature was scaled by its standard deviation obtained from the variants in the training set.

3.3.3. Construction of the model

We assigned class label 0 to the proxy benign/neutral variants and 1 to the proxy deleterious variants. Then, we trained a logistic regression classifier to predict the posterior probability of a variant being proxy deleterious. We used the logistic regression module provided by Graphlab v2.1 [37]. Based on previous experience and given the lack of a sufficiently large validation set, we applied the set of hyper parameters that were found to be optimal for mouse CADD19, i.e. L2-penalization was set to 0.1 and the number of iterations to 100. Feature rescaling, performed by the logistic regression function by default, was deactivated.

3.3.4. Score creation

The pCADD scores were computed for all potential SNVs (3 per position) on the 18 autosomes and the X allosome. Each SNV was annotated with 867 genomic annotations and scored by the trained logistic regression model. Subsequently, these scores were sorted in descending order and assigned a pCADD score defined as $-10 * \log_{10}(\mathrm{i/N})$, with i being the rank of a particular SNV and N the total number of substitutions (N = 7,158,434,598).

3.3.5. Analyses

3.3.5.1. Codon analysis

From the Ensembl v.93 pig gene build, we retrieved 10,942 genes with only one annotated transcript to avoid complications due to overlapping transcripts. We created three sets, consisting of the minimum pCADD score found at a site, per transcript, one for each of the three positions of a codon. We computed one-tailed Mann–Whitney U-tests between each of the three sets. The resulting p-values were Bonferroni corrected. All calculations were performed in Python version 3 using SciPy v.1.1.0 [38] and Statsmodels v.0.9.0 [39].

3.3.5.2. *miRNA analysis*

We obtained all annotated (pre-)miRNA sequences from the Ensembl v93 database, i.e. 484 sequences, and, after removal of sequences that overlapped with any of the training SNVs, 294 sequences remained. As a second set, equally long sequences up- and downstream of the miRNA sequence were selected. For each position in both sets, the miRNA sequences and surrounding sequences were annotated with the maximum pCADD score. To test whether miRNA sequences had a significantly higher pCADD score than their neighbouring sequences, we applied a one-tailed Mann–Whitney U-test using SciPy v.1.1.0 in Python 3.

3.3.5.3. *Intron analysis*

We used the REST API of Ensembl v93 to download the intron coordinates of all 40,092 transcripts. We annotated all the sites in all the introns with the maximum pCADD score found at these sites. For each intron, we performed one-tailed Mann–Whitney U-tests to check if the investigated intron had a significantly higher pCADD score than all the other introns in the same transcript. p-values were Bonferroni corrected over all transcripts, per intron. To display the results, we normalized the number of rejected null-hypotheses by the number of conducted tests, which decreases as the number of introns increases.

3.3.5.4. *Tissue analysis*

We downloaded porcine Affymetrix expression data of several tissues published by Freeman et al. [40]. We selected the genes that were clustered and associated with a particular tissue and had a robust multi-array average (RMA) [41] expression level of at least 100 or more to filter out genes with no activity. Of these genes, we considered all the coding DNA sequences (CDS); if a particular CDS was present in more than one transcript, it was selected only once. In addition to the housekeeping genes, genes specific for 16 tissues were selected (cartilage-tendon, blood, cerebellum, dermal, epithelium, eye, kidney, liver, lung, muscle, neurone, pancreas, placenta, salivary gland, testis, and vasculature). All CDS were annotated with the maximum pCADD score found at each site of the CDS and merged into one set per tissue. Tissue sets were tested for higher scores than those of the housekeeping set with one-tailed Mann–Whitney U-tests; p-values were Bonferroni corrected. All calculations were done in Python 3 using the SciPy v.1.1.0 and Statsmodels v.0.9.0. modules.

3.4. Results

In this study, we trained a CADD-like model for SNV prioritisation in the pig genome, which is referred to as pCADD. It is a linear regressor that is trained to differentiate between two classes of variants, a set of simulated variants, which is relatively more enriched in potentially deleterious variants than a set of derived variants, which is depleted in deleterious variants. The pCADD generated a score for every possible SNV of the Sscrofa11.1 reference genome on all autosomes and the X allosome. Then, these scores were tested on a held-out test set, they were used to evaluate seven SNVs with known functional effect and we examined whether they could discriminate between functional and non-functional sequences.

3.4.1. pCADD data characteristics

The class distribution in the training and test sets were balanced, but subsets of SNVs found in different genomic regions displayed varying proportions of simulated and derived SNVs (Table 1). These imbalances were similar to those found for the human (hCADD) and mouse (mCADD) datasets in our previous study [19]. The largest difference among the three models is the total number of SNVs used for model training: ~31 million for hCADD, ~67 million for mCADD and ~112

million for pCADD. This results from the use of a more distant ancestor of the pig than the ancestors used for mouse in mCADD (mouse and rat) and for humans in hCADD (human and chimpanzee). A more distant ancestor yields more differences between the inferred ancestor and the species of interest, resulting in a larger derived class and, thus, in a larger total number of SNVs to create a balanced dataset.

Table 1: The number of SNVs and their relative proportions of the six subsets of the test set for pCADD.

Pig partition	Number SNVs / proportion of test set	Num. Simulated	Num. Derived	Class distribution (Simulated/Derived)
Test set	11,197,628 / 100.00%	5,598,814	5,598,814	50.00% / 50.00%
Not cDNA	10,884,147 / 97.20%	5,404,059	5,480,088	49.65% / 50.35%
cDNA	313,481 / 2.80%	194,755	118,726	62.13% / 37.87%
Not CDS	154,622 / 1.38%	84,730	69,892	54.80% / 45.20%
CDS	158,859 / 1.42%	110,025	48,834	69.26% / 30.74%
Synonymous	75,216 / 0.67%	40,147	35,069	53.38% / 46.62%
Missense	83,643 / 0.75%	69,878	13,765	83.54% / 16.46%

3.4.2. Increased discriminative power of pCADD with increased biological relevance of the sequence in which the queried SNVs are located

The performance of pCADD is evaluated by computing the receiver-operator-area under the curve characteristic (ROC-AUC) on a test set, which consisted of simulated and derived SNVs, none of which were used for training. The overall ROC-AUC on the entire test set is ~0.683 but differs considerably for six subsets of SNVs (Figure 2a). The test sets are subsets of each other, with decreasing numbers of SNVs beginning with the whole test set and ending with the missense mutations. In transcribed regions of the genome, the scores are more discriminative than in non-transcribed regions, while in coding regions they are more discriminative than in non-coding regions such as the 5'UTR, 3'UTR and introns. The scores are most discriminative for missense mutations, which have the largest number of genomic annotations, resulting in high discriminative performance of the pCADD model.

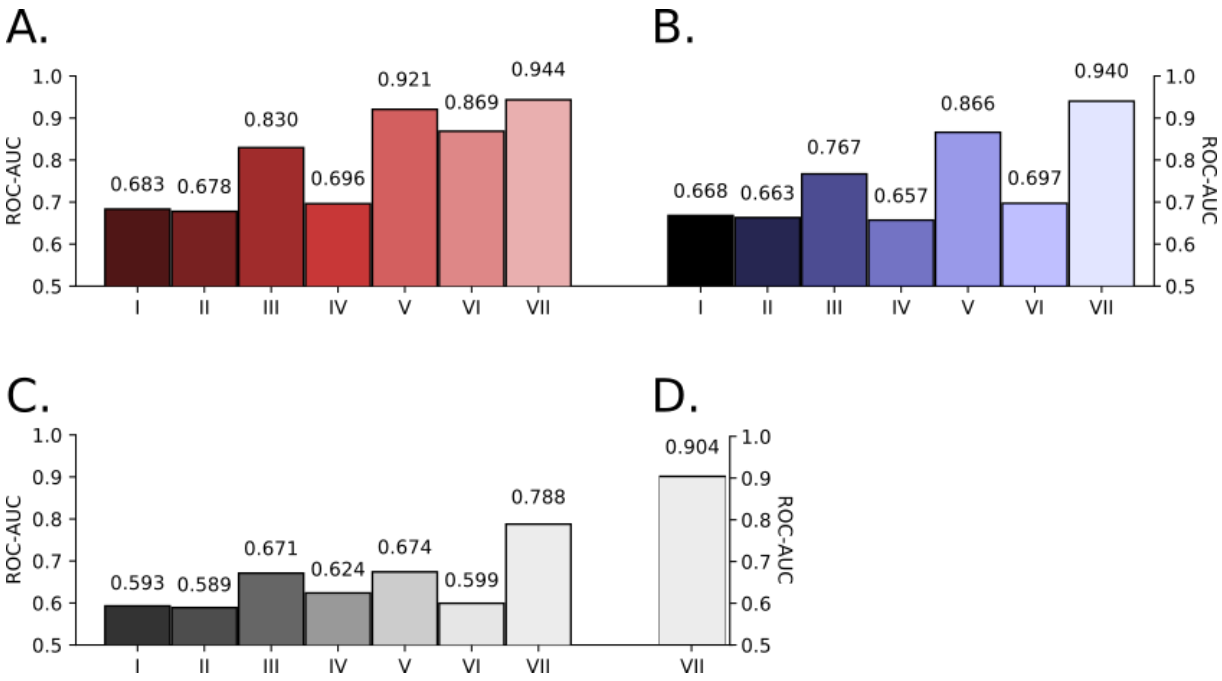


Figure 2: This figure displays the prediction performances of different prioritization tools on test sets, representing different regions of the genome for which different number of features are available. I: Whole test set; II: Intergenic SNVs; III: Transcribed SNVs; IV: SNVs in intron, 5' & 3' UTRs; V: Coding SNVs; VI: SNVs causing synonymous mutations; VII SNVs causing missense mutations. A) pCADD performance measured in ROC-AUC on the different subsets of the pig held-out test set. B) mCADD test performance measured in ROC-AUC on the same genomic subsets in the mouse genome. C) Performance of 6-taxa laurasiatheria PhastCons conservation score on the pig test set. D) SIFT performance on missense causing SNVs in the pig test set.

These observations are in strong accordance with the earlier reported observations for the mCADD model for mouse (reproduced in Figure 2b) [19], which was proven useful to identify truly deleterious mutations found in the Mutagenetix [42] data base, lifted from ClinVar [43] and others [19]. For all investigated SNV subsets, PhastCons [28] conservation scores based on the Ensembl 6-taxa laurasiatheria [32] displayed the same pattern across all subsets, but performed worse than pCADD (Figure 2c). We used 6-taxa laurasiatheria PhastCons scores because, overall, they performed best on different subsets of the held-out test set (see Additional file 3: Figure S1). A similar difference in performance was observed when the performance of pCADD on missense mutations was compared to that of SIFT (Figure 2d), which indicates the added value of pCADD over conventional approaches of identifying potential candidates.

3.4.3. Selecting candidate SNVs based on their total score and on their relative rank in the surrounding region is meaningful

When we assessed examples of known causal SNVs (Table 2), they were enriched in the upper percentile of pCADD scores and were likely to be picked up as potential. The exception is 3:43952776T>G, one of two variants located in close proximity to a splice-site. In particular, it is located in an intron sequence, 4 bp upstream of an annotated splice site. Variants, which are located 1- and 2-bp upstream of the splice site have pCADD scores that range from 20.90 to 21.93, whereas the remaining variants in the same intron sequence have on average a pCADD score of ~2.96. Only 13 (out of 3450) other potential SNVs in that intron have a higher pCADD score. This puts the 3:43952776T>G SNV into the 99.6th percentile of the intron sequence in

Table 2: Five well-known examples of causal SNVs with different effects on phenotype and their pCADD scores. The pCADD scores and percentiles both indicate their rank among all potential SNVs in the pig genome.

Genomic location	Substitution	pCADD	Percentile	Gene	Effect	Citation
6:146829589	G>A	22.868	99.5	LEPR	missense: affects productive, fatness and meat quality traits in different genetic backgrounds	[44]
1:265347265	A>G	17.198	98.1	NR6A1	missense: affects number vertebrae	[45]
17:57932233	A>C	23.322	99.5	PCK1	missense: causal mutation associated to intramuscular fat content, backfat thickness and meat quality in pigs	[46]
7:31281804	G>A	21.589	99.3	PPARD	missense: affects ear size, fat metabolism, skin and cartilage development	[47]
12:38922102	G>A	21.848	99.3	TADA2A	splice-donor: lethal recessives	[48]
3:43952776	T>G	10.144	90.3	POLR1B	splice-region: lethal recessives	[48]
6:54880241	T>C	28.767	99.9	PNKP	missense: lethal recessives	[48]

which it is located. None of the 13 potentially higher scored variants were observed in our population of 384 pigs, which makes 3:43952776T>G the highest scored SNV in that region.

3.4.4. The third position of a codon is scored lower than the first two

To assess further if the model assigns different scores to sites with differing biological importance genome-wide, we tested whether the three positions in a codon are scored differently. Based on the fraction of non-synonymous mutations for each codon position, the second position should receive the highest score, followed by the first and third positions (see Additional file 3: Figure S2). To test this, we examined codons of genes that have only one known transcript, to avoid interference, which is expected by overlapping transcripts.

The table displays the counts of significant p-values between the three different positions in a codon. The columns indicate the positions that are tested to have higher pCADD scores than the positions in the rows. The numbers indicate how often the null hypothesis was rejected in 10,942 conducted tests.

Table 3: Bonferroni corrected one tailed Mann-Whitney U tests were conducted to test if pCADD values are significantly larger in one codon position relative to another. The table displays the counts of significant p-values between the three different positions in a codon. The columns indicate the positions that are tested to be larger than the positions in the rows. The numbers indicate how often the null hypothesis was rejected in 10,942 conducted tests.

Smaller \ larger	1st	2nd	3rd
1st	NA	3066	189
2nd	766	NA	340
3rd	8830	8901	NA

Table 3 shows the number of significant tests when comparing the pCADD scores between two codon positions, across a gene, with each other (Bonferroni corrected, one-tailed Mann–Whitney U-tests). Among the 10,942 genes that were selected for this test, we found that the second codon position has a significantly higher pCADD score than the third for 8901 genes, and that the first codon position has a significantly higher pCADD score than the third for 8830 genes. Only for 3066

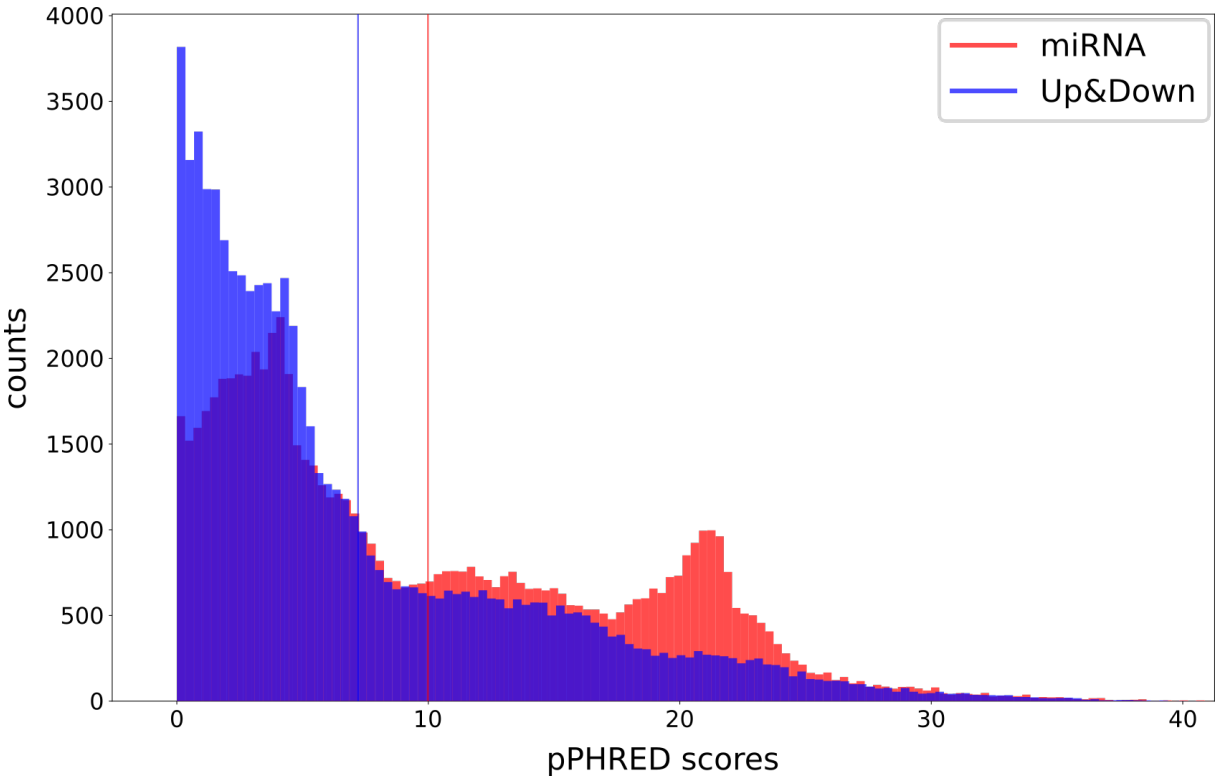


Figure 6: Histogram of pCADD score distribution of (pre-)miRNA transcripts and their surrounding up- and downstream regions. Vertical lines indicate the mean values of each distribution with a mean of 9.987 for miRNA and 7.205 for Up&Down. The One-tailed Mann-Whitney U-test between both distributions returned a p-value of 0.0 and a ROC-AUC of 0.613 in favour of miRNA over the Up&Down stream regions.

genes, did the second codon position score significantly higher than the first, while for 766 genes it was the opposite. Taken together, these results agree with our expectation, and indicate that pCADD scores do reflect deleteriousness. This was further confirmed by comparing the effect sizes, measured as ROC-AUC of the pairwise comparisons of codon positions (see Additional file 3: Figure S3).

3.4.5. miRNA regions are scored differently from those of neighbouring regions

We investigated whether pCADD scores are higher for functional non-coding sequences than for non-functional sequences up- and downstream. Variants in annotated (pre-)miRNA regions have significantly higher pCADD scores (p-value=0.0, one-tailed Mann–Whitney U test; ROC-AUC=0.613) than sites in up- and downstream regions (average pCADD scores of ~10 vs. ~7.2) (Figure 3). This difference is largely due to an abundance of (pre-)miRNAs with pCADD scores around ~21 and a relatively smaller number of variants with a low score. For 164 miRNAs (~56%), the pCADD scores were significantly higher than those of the neighbouring regions (Bonferroni corrected, one-tailed Mann–Whitney U test).

3.4.6. Among the introns of a transcript, the first one has the highest score

Chorev et al. [49] showed that regulatory elements are enriched in the first few introns of a transcript and that their number decreases with increasing intron position. Consequently, we expected to see decreasing pCADD scores with increasing intron position. To test this, we annotated every position in the intron region with the highest pCADD score for that position and

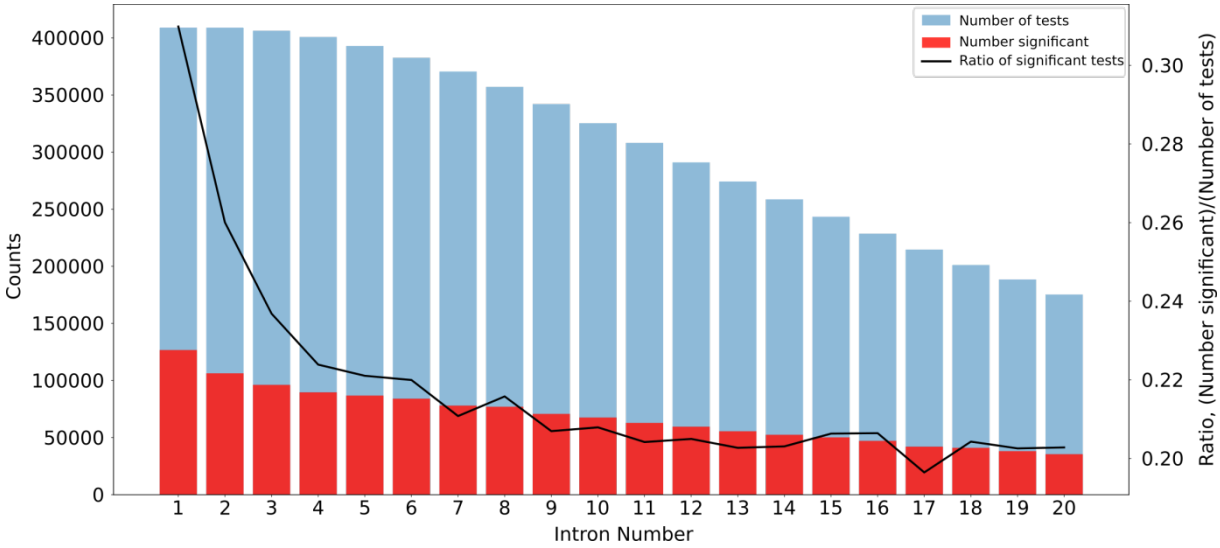


Figure 7: pCADD scores per intron compared to all other introns, for the first 20 introns. The blue bar indicates the number of introns tested against the intron of interest, the red bar shows how many of these tests resulted in an adjusted p-value < 0.05 (scale on the left axis). With increasing intron position, the number of tests that can be conducted decreases (with the number of transcripts that have at least that many introns). The black line represents the normalised number of significantly enriched introns, normalized by the number of conducted tests per intron position (scale on the right axis).

calculated how often the scores in a particular intron are significantly higher than those across all other introns in the same transcript (Bonferroni corrected one-tailed Mann–Whitney U test). The results clearly show that introns closer to the transcription start site of a gene have higher pCADD scores (Figure 4), which provide evidence for their biological relevance.

3.4.7. Among all tested tissues, pCADD scores for salivary glands and neuronal tissue specific genes are the lowest and highest, respectively.

Next, we investigated whether genes considered to be housekeeping genes have different (higher) pCADD scores than genes specifically expressed in certain tissues. The underlying assumption is that a mutation in a gene expressed in all tissue types has a much broader potential deleterious effect. We compared pCADD and PhyloP scores of genes specific for 16 tissues and also compared them (Bonferroni corrected one-tailed Mann–Whitney U test; ROC-AUC) to scores of a set of genes considered as housekeeping genes, i.e. expressed approximately equally in all tissues [40]. Based on pCADD scores, housekeeping genes had significantly higher scores for 12 of the 16 tissues examined (Table 4). Genes in three brain-derived tissues—cerebellum, eye, neuronal tissue—and in muscle tissue (smooth and skeletal) have on average a higher pCADD score than housekeeping genes. A ROC-AUC of 0.5 is the expected performance if the pCADD scores are randomly assigned to the genes of each set. This means that the larger the absolute difference is from 0.5, the clearer is the signal supporting that one set is larger than the other. We compared all tissue gene sets to housekeeping genes, this means that when the ROC-AUC is smaller than 0.5, the pCADD scores of the tissue associated gene set are generally larger than those of the housekeeping one and vice versa. In all the comparisons, the total effect size was small and did not differ from 0.5 by more than 0.122 (dermal tissue). The four tissues that displayed higher pCADD scores than housekeeping genes have in common that their cells do not divide anymore once they are fully differentiated. Mutations in these tissues may have a larger effect than in tissues with a high rate of cell division due to the inability of the tissue to replace cells, which leads to scarring and

eventually tissue failure. Thus, genes specific to these four tissues are more likely conserved than those specific to other tissues, resulting in overall higher pCADD scores. This is supported by the analysis with conservation scores (Table 4), which showed that these genes were more conserved than the housekeeping genes. Tissues such as dermal and salivary gland show the lowest pCADD scores and high rates of cell division. These tissues are likely more tolerant to germline mutations since they must adapt to changes in diet and climate, thus their tissue-specific genes have a higher variability, resulting in lower pCADD scores.

Table 4: Test results between tissue specific gene sets and house-keeping genes. We tested if tissue specific genes are significantly lower scored than house-keeping, using pCADD and PhyloP scores (25-taxa mammalian alignment). The ROC-AUC scores display the likelihood that a random sample from the scores of the house-keeping genes is larger than one from the scores of tissue specific genes.

Tissue	pCADD *p*-value (tissue < house-keeping)	pCADD ROC-AUC (house-keeping vs tissue)	PhyloP p-value (tissue < house-keeping)	PhyloP ROC-AUC (house-keeping vs tissue)
All tissues	2×10^{-1}	0.5	1	0.467
Blood	3×10^{-122}	0.512	1	0.481
Cartilidge-tendon	3×10^{-35}	0.511	1	0.453
Cerebellum	1	0.48	1	0.487
Dermal	0	0.622	0	0.681
Epithelial	0	0.538	1×10^{-29}	0.515
Eye	1	0.475	1	0.456
Kidney	2×10^{-100}	0.515	1	0.468
Liver	1×10^{-54}	0.51	9×10^{-1}	0.49
Lung	6×10^{-8}	0.506	1×10^{-2}	0.503
Musculature	1	0.491	1	0.468
Neuronal	1	0.443	1	0.4
Pancreas	1×10^{-310}	0.558	$3x10^{-81}$	0.559
Placenta	1×10^{-145}	0.529	1	0.469
Salivary-gland	7×10^{-48}	0.519	1	0.478
Testis	0	0.558	1	0.478
Vasculature	0	0.558	1	0.454

3.4.8. Differentiation between functional and non-functional sequences is greater with pCADD than conservation scores

Conservation scores are often used to evaluate the potential importance of sequences and to evaluate if a particular candidate SNV may have a deleterious effect. They are also useful to put our own results into perspective and assess conventional sequence prioritisation methods.

Similar to the section "miRNA regions are scored differently from those of neighbouring regions", we annotated the pre-miRNAs and their associated up- and downstream regions with PhyloP

conservation scores (based on 25-taxa mammalian alignment) and performed the same analysis by computing significance tests to check if miRNA sequences have higher pCADD scores than those in their neighbouring regions. We chose 25-taxa PhyloP scores because these have the largest coverage of the pig genome among all conservation scores used in this study (see Additional file 2: Table S4). The results are in Additional file 3: Figure S4 and are very similar to those from the analysis using pCADD scores, with an almost identical p-value close to 0 (1e−225) and a ROC-AUC value of 0.595, which indicates a slightly worse separation between both classes of sequences than when using pCADD.

Likewise, we evaluated the intron positions relative to each other using the same PhyloP conservation scores to annotate intron sequences. The results in Additional file 3: Figure S5 show a similar pattern of decreasing importance with increasing intron position as observed when the introns are annotated with pCADD scores. Major differences between the analysis using pCADD and conservation scores is that the total number of introns, which can be annotated with conservation scores is smaller, resulting in 81,743 fewer tests compared with pCADD. Furthermore, the ratio between the total number of tests and the number of tests with an adjusted significant p-value is smaller when conservation scores are used, which indicates that conservation scores are less discriminative between different intron positions.

We annotated tissue-specific and housekeeping genes with PhyloP conservation scores to investigate whether the differentiation between both sets of genic regions followed the same pattern. Twelve tissue-specific gene sets displayed significantly lower pCADD scores than housekeeping genes, whereas only four tissues had a significantly lower conservation score. The larger total differences in ROC-AUC scores obtained by using PhyloP scores compared to pCADD scores indicate that the variations between tissue gene sets are larger when using PhyloP.

The worse performance of PhyloP scores to distinguish between pre-miRNA and surrounding regions is supported by the lower ratio of significant tests in the intron analysis, which indicates that PhyloP scores have less specificity for functional elements than pCADD scores.

3.4.9. Predicted intergenic SNVs with high pCADD scores are often associated with lncRNA and may indicate missing annotations

To examine the utility of pCADD scores for the prioritization of SNVs, we investigated whether they can help in the identification of intergenic candidate SNVs that segregate between two closely related Large White pig breeding populations. We scored intergenic SNVs that were unique for either of these pig populations by multiplying their pCADD score with the allele frequency and selected the top 20 highest scored SNVs for each population. Since the pCADD model is based on the Ensembl pig annotations [50] (Ensembl gene annotation update e!90 Sscrofa11.1), we matched the selected 40 SNVs with NCBI's pig gene build [51] to determine whether the model captures non-annotated genomic features. We found that 16 of the 40 SNVs are located within a (NCBI) coding region (one example shown in Figure 5) and six SNVs overlap with a (NCBI) long non-coding RNA (Table 5).

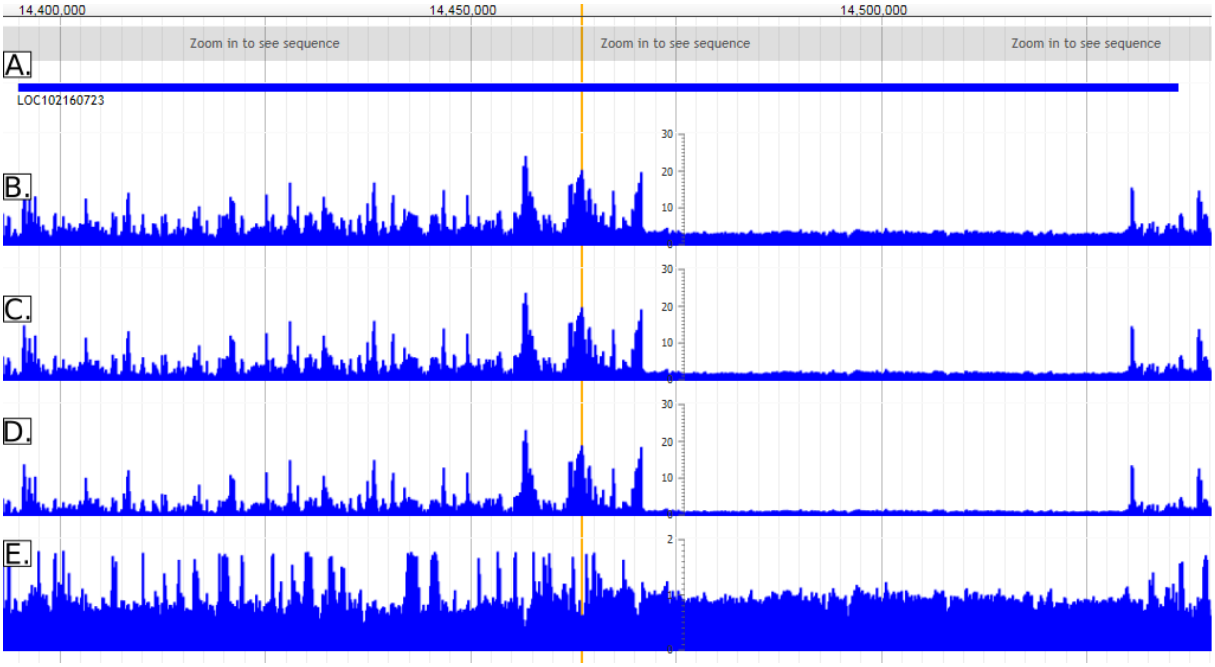


Figure 8: There are three different potential nucleotide substitutions at each position in the genome, each with their own predicted pCADD score. To visualize them in JBrowser [52] we created tracks for the maximum, median and minimum scores at each position. The fourth track is displaying the standard deviation among the three scores to identify more easily sites of variable deleteriousness. The yellow vertical bar is located at position 5:14463457, indicating the site of the top scoring SNV in Table 5. This SNV is considered intergenic according to the Ensembl gene build but located within a lncRNA according to the NCBI genebuild. A) NCBI gene build track, showing the genomic region belonging to lncRNA LOC102160723. B,C,D) the maximum, median and minimum pCADD scores for each position in the displayed region. E) The standard deviation of pCADD scores at each position.

Table 5: Top 40 SNVs according to pCADD*Alt:Frq, which are presumably intergenic according to the Ensembl Sus scrofa gene build, annotated with NCBI. When no NCBI gene annotation was found they were mapped to hg38 and the Human Ensembl gene build was used. Blue: SNVs that are intergenic in the three gene builds, yet found in regions with conserved synteny. Red: SNVs located in a region unannotated in any gene build.

Chr	Pos	Ref:Frq	Alt:Frq	pCADD	pCADD*Alt:Frq	NCBI-gene build	Human-Ensembl-gene build
5	14463457	T:0.014	C:0.986	26.559	26.185	lncRNA	
10	45490687	G:0.007	T:0.993	24.175	24.000	RSU1	
9	88698813	C:0.021	G:0.979	24.433	23.909		lncRNA
6	149549021	T:0.007	C:0.993	23.714	23.544		
18	30883512	G:0.045	A:0.955	24.211	23.111	lncRNA	
14	102653354	A:0.007	G:0.993	23.216	23.052	lncRNA	
3	35533299	C:0.029	T:0.971	23.729	23.041	RBFOX1	
8	16080284	T:0.021	G:0.979	23.540	23.035	KCNIP4	
8	16090742	A:0.007	C:0.993	23.188	23.0248	KCNIP4	
9	88631400	T:0.037	C:0.963	23.855	22.978		lncRNA
13	11996804	A:0.068	G:0.932	24.518	22.846		miscRNA
8	16069085	C:0.014	T:0.986	23.148	22.817	KCNIP4	
1	270976051	G:0.057	A:0.943	24.148	22.768		
12	10080096	C:0.029	T:0.971	23.417	22.738		
15	134154371	G:0.028	A:0.972	23.388	22.729		
17	15317464	T:0.035	C:0.965	23.437	22.611		
8	16126909	T:0.145	G:0.855	26.331	22.515	KCNIP4	
14	102708028	T:0.007	C:0.993	22.622	22.463		lncRNA
17	8460314	T:0.007	A:0.993	22.607	22.448		FAT1
3	2721065	C:0.016	T:0.984	22.794	22.438		SDK1
8	2274651	T:0.006	C:0.994	24.861	24.721	lncRNA	
14	41547002	T:0.006	C:0.994	24.651	24.511	MYO1H	
9	88656584	T:0.023	C:0.977	24.606	24.047		lncRNA
13	145274213	A:0.031	G:0.969	24.336	23.576	ZBTB20	
5	14463352	A:0.006	G:0.994	23.526	23.393	lncRNA	
2	135162568	A:0.011	C:0.989	23.305	23.043		
13	196634107	A:0.011	C:0.989	23.190	22.930		lncRNA
13	203405436	G:0.006	A:0.994	23.046	22.917		
17	15317464	T:0.022	C:0.978	23.436	22.910		
13	203404345	T:0.017	G:0.983	23.239	22.842		
18	4227731	C:0.006	A:0.994	22.839	22.710		
13	203405428	T:0.006	G:0.994	22.663	22.535		
13	145279451	A:0.019	G:0.981	22.960	22.512	ZBTB20	
15	134347171	T:0.006	G:0.994	22.633	22.506		
5	25295998	A:0.011	G:0.989	22.731	22.476	lncRNA	
15	134154371	G:0.040	A:0.960	23.387	22.457		
18	42017803	T:0.017	G:0.983	22.811	22.427		
15	134347189	G:0.006	C:0.994	22.471	22.345		
8	16126909	T:0.152	G:0.848	26.331	22.337	KCNIP4	
14	138794865	A:0.006	G:0.994	22.411	22.285		lncRNA

In addition, we mapped the genomic locations of the candidate SNVs to the human assembly GRCh38.p12 and Ensembl gene builds, which revealed nine additional genic regions that consisted of six lncRNAs, one region considered as a miscRNA and two genes. For all 40 SNVs, synteny of the surrounding genes was conserved except for 18:4227731C>A. The relatively large number of prioritized SNVs that overlap with lncRNAs can be explained in two ways. First, there might be a considerable number of missing annotations in the gene builds that we used because the RNA-seq

databases are incomplete and are the basis for lncRNA annotations. Second, although the lncRNA functions are conserved due to islands of strong conserved regions [53], the architecture of their sequences experience constant restructuring and weak sequence conservation across species [53], [54].

The highest scored SNVs (in terms of pCADD score multiplied by alternative allele frequency) for which no genic annotation was found (6:149549021T>C) (Table 5), is located in an island with high pCADD scores within a region that contains several of such small islands (see Additional file 3: Figure S6). This region starts with a highly H3K27Ac acetylated region, which indicates an enhancer site. Such a pattern is uncommon for intergenic regions and could indicate a missing annotation in the gene builds used in our study.

3.5. Discussion

We used a method that provides scores for the prioritization of SNVs with respect to their putative deleteriousness, from which we derived functional relevance for the genomes of pig. The method is based on the creation of a set of derived variants from an inferred common ancestor sequence that can be assumed to be depleted in deleterious variants and a set of simulated variants that are likely to be enriched in variants with a deleterious effect. It is important to note that while it is reasonable to assume that the proxy benign/neutral are truly benign/neutral variants, the simulated putative deleterious variants may also encompass a relatively large proportion of actually neutral variants.

Founder effects in pig populations may lead to the accumulation of functional variants, with both benign and deleterious variants receiving a relatively high pCADD score. This means that pCADD scores are useful to prioritize SNVs of interest, but that assessing deleteriousness may need additional information or experiments. For example, the missense variant 1:265347265A>G (pCADD:21.848), which is responsible for an increased number of vertebrae and can be considered benign given current breeding goals, and the deleterious lethal recessive splice variant 12:38922102G>A, have similar pCADD scores (pCADD: 17.198) (Table 2).

We evaluated the generated pCADD scores on a held-out test set and reported performances on different genomic subsets, which we compared to results of our previous study on mouse. Due to the nature of the procedure, the test performance can only indicate if the training algorithm has picked up patterns of features that are predictive for the simulated variants and if the performance varies with the genomic region. It has to be emphasized that only performance trends can be meaningfully compared between the different mCADD/pCADD models due to the different datasets used for computation. In spite of the large number of neutral variants, which is expected in both sets of variants, the performance seems to indicate that patterns to differentiate between the derived and simulated datasets have been picked up and can be used to evaluate variants and regions based on their potential interest.

The performance of pCADD scores to discriminate between simulated and derived variants in the test set increased as the number of features increased, depending on the genomic regions in which they are embedded. The consequence is that missense mutations are the best classified, although the most interesting application of pCADD is to annotate non-coding and intergenic variants, for which a plethora of functional candidates exist but there are only a few methods for further prioritization. As shown for the splice-region variant 3:43952776T>G, the ranking of a variant relative to its neighbouring sequence in the same sequence category (introns, exons, intergenic, etc.) can provide information that helps to prioritize such variants.

Furthermore, we used PHRED-like scores to rate different sequences with known biological function. We compared the scores for the three positions in a codon and found that less redundant positions achieve higher pCADD scores. Moreover, regulatory sequences could be clearly distinguished from their neighbouring regions (i.e. high scores in miRNAs). In addition, our model supports the higher frequency of regulatory elements in the first few introns of a transcript, and thus has the potential of scoring not only individual SNVs but also of using a summary score per site to annotate entire regions to identify potential sub-regions of interest. This is a clear advantage compared to alternative methods to evaluate non-coding sequences, such as conservation scores, which may not be available for the entirety of the genome. This was the case in the analysis of intron sequences, for which more than 80,000 fewer tests could be conducted due to missing conservation scores. Using pCADD, candidate regions in which annotations are potentially missing can be identified. For example, no annotation was found for the 6:149549021T>C SNV, even though pCADD scores were within a range typical for exons and displayed patterns of islands of high importance (see Additional file 3: Figure S6), which is more compatible with coding regions than with intergenic regions. Ensembl gene annotations rely strongly on transcript data from public databases, which implies that incomplete databases may lead to missing gene annotations. This is especially the case for species that are less well studied than model organisms or humans. In addition, if the genes in question are not ubiquitously expressed, they can be absent from the data of the sequenced tissue. The same is true for genes, the expression of which depends on developmental-, disease- or physiological state, as is the case for many lncRNAs [54].

We compared genes specific for 16 different tissues against (presumed) housekeeping genes [40]. Our assumption was that the ubiquitously and generally more highly expressed housekeeping genes [55] should have globally higher scores than tissue-specific genes. Although the absolute effect size was small, significantly higher scores were attributed to genes specific to cerebellum, eye, neuronal and muscle tissue. Brain-derived tissues (cerebellum, eye, neuronal tissue), in particular, displayed the largest effect sizes. On the one hand, brain tissue has experienced major development changes during the time period between 535 and 310 Mya ago, i.e. increased expression and gain of functions of paralogs of brain-specific genes [56], [57]. Since then and during the entire mammalian development, the expression of paralogs of brain-specific genes is lower than that observed in other tissues [57], which indicates the fine balancing that acts to keep the brain functional. This emphasizes the extreme importance of brain-specific genes for survival and probably their low tolerance to mutations, compared to housekeeping genes. On the other hand, dermal tissue (epithelium) is one of the most ancient tissues in the evolution of metazoans and has highly conserved developmental pathways, which include genes that are involved in the adaptation to specific environmental changes and have overall lower pCADD scores than housekeeping genes.

Among the most important features for the pCADD model are conservation scores. They are annotated for large fractions of the genome (see Additional file 2: Table S4), and thus they heavily influence training. This is supported by our investigation of various tissues, which showed that particularly high scores were assigned to expected strongly conserved regions. Deleterious effects that are not captured by sequence conservation, such as changes in the epigenome or in relatively variable regions, are expected to have lower scores. This becomes problematic when the species of interest has experienced recent genetic bottlenecks and has been subjected to very strong selection, which change the species' genotype, as is the case for domesticated species. In this case, the patterns observed from evolutionary changes may not be accurate to evaluate recent changes. However, not all the regions in the genome are subject to substitution, neither in natural nor in domesticated environments. There are exceptions to this rule, such as the reported missense mutations in Table 2, which are causal for a change in the number of vertebrae, ear size,

meat quality and fat content, and have high scores, which support the use of pCADD for variant prioritization.

3.6. Conclusions

The CADD approach is widely used in humans [13]–[18] and, based on our findings, it seems to be a suitable approach for pig (and other non-human species). Variants that distinguish populations can be ranked with respect to their pCADD score and allele frequency to find potential candidates for phenotypes expressed in the studied populations. pCADD could become a valuable tool in pig breeding and conservation. It can be used to score variants with a potential negative effect in small-sized endangered local pig breeds, but also help prioritize high-impact variants in genomic prediction to further enhance genomic selection.

3.7. Declarations / Statements

3.7.1. Availability of data and material

pCADD scores, partitioned per chromosome, compressed via bgzip and tabix indexed for fast access, can be downloaded following this link (~5–1 GB): http://www.bioinformatics.nl/pCADD/indexed_pPHRED-scores/

To create tracks for genome browsers we provide the maximum, median, minimum, and standard deviation summaries of each site, partitioned per chromosome. All files are compressed with bgzip and tabix indexed and can be downloaded following this link (~1.7 GB to ~350mb): http://www.bioinformatics.nl/pCADD/indexed_pPHRED-summary-scores/

Scripts and data to recreate the figures in this article can be downloaded from the following link: https://git.wur.nl/gross016/pcadd-scripts-data

3.7.2. Funding

This research was funded by the TTW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality. This study was financially supported by NWO-TTW and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs-Norsvin. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Bibliography

[1] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[2] B. H. Good and M. M. Desai, "Deleterious passengers in adapting populations," *Genetics*, vol. 198, no. 3, pp. 1183–1208, 2014.

[3] J. H. Gillespie, "Is the population size of a species relevant to its evolution?," *Evolution*, vol. 55, no. 11, pp. 2161–2169, 2001.

[4] M. Pérez-Enciso, J. C. Rincón, and A. Legarra, "Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised," *Genet. Sel. Evol.*, vol. 47, no. 1, pp. 1–14, 2015.

[5] R. F. Brøndum *et al.*, "Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction," *J. Dairy Sci.*, vol. 98, no. 6, pp. 4107–4116, 2015.

[6] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Res.*, vol. 11, no. 5, pp. 863–874, 2001.

[7] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[8] M. Hecht, Y. Bromberg, and B. Rost, "Better prediction of functional effects for sequence variants," *BMC Genomics*, vol. 16, no. 8, p. S1, 2015.

[9] Y. Choi and A. P. Chan, "PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics*, vol. 31, no. 16, pp. 2745–2747, 2015.

[10] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: Predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, 2019.

[11] B. Guiko, M. J. Hubisz, I. Gronau, and A. Siepel, "Probabilities of fitness consequences for point mutations across the human genome," *Nat. Genet.*, vol. 47, no. 3, pp. 276–283, 2015.

[12] Y. F. Huang, B. Gulko, and A. Siepel, "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data," *Nat. Genet.*, vol. 49, no. 4, pp. 618–624, 2017.

[13] H. E. Abboud *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2017.

[14] J. K. van der Velde *et al.*, "Evaluation of CADD Scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization," *Hum. Mutat.*, vol. 36, no. 7, pp. 712–719, 2015.

[15] S. Balasubramanian *et al.*, "Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes," *Nat. Commun.*, vol. 8, no. 1, 2017.

[16] B. Banaganapalli *et al.*, "Comprehensive computational analysis of GWAS loci identifies CCR2 as a candidate gene for Celiac disease pathogenesis," *J. Cell. Biochem.*, vol. 118, no. 8, pp. 2193–2207, 2017.

[17] M. Mesbah-Uddin, R. Elango, B. Banaganapalli, N. A. Shaik, and F. A. Al-Abbasi, "In-silico analysis of inflammatory bowel disease (IBD) GWAS loci to novel connections," *PLoS One*, vol. 10, no. 3, pp. 1–21, 2015.

[18] N. A. Al-Tassan *et al.*, "A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer," *Sci. Rep.*, vol. 5, no. January, pp. 1–11, 2015.

[19] C. Groß, D. de Ridder, and M. Reinders, "Predicting variant deleteriousness in non-human species: Applying the CADD approach in mouse," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, 2018.

[20] M. A. M. Groenen *et al.*, "Analyses of pig genomes provide insight into porcine demography and evolution," *Nature*, vol. 491, no. 7424, pp. 393–398, 2012.

[21] A. V Zimin *et al.*, "A whole-genome assembly of the domestic cow , Bos taurus," *Genome Biol.*, vol. 10, no. 42, 2009.

[22] Y. Jiang *et al.*, "The sheep genome illuminates biology of the rumen and lipid metabolism," *Science*, vol. 344, no. 6188, 2014.

[23] B. Paten, J. Herrero, K. Beal, S. Fitzgerald, and E. Birney, "Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs," *Genome Res.*, vol. 18, no. 11, pp. 1814–1828, 2008.

[24] B. Paten *et al.*, "Genome-wide nucleotide-level mammalian ancestor reconstruction," *Genome Res.*, vol. 18, no. 11, pp. 1829–1843, 2008.

[25] S. W. Doniger *et al.*, "A catalog of neutral and deleterious polymorphism in Yeast," vol. 4, no. 8, 2008.

[26] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor.," *Genome Biol.*, vol. 17, no. 1, p. 122, 2016.

[27] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, 2010.

[28] A. Siepel *et al.*, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.," *Genome Res.*, vol. 15, no. 8, pp. 1034–50, 2005.

[29] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Comput. Biol.*, vol. 6, no. 12, 2010.

[30] G. R., "Amino acid difference formula to help explain protein evolution.," *Science*, vol. 185, no. 4154, pp. 862–4, 1974.

[31] T. Zhou *et al.*, "DNAshape : a method for the high-throughput prediction of DNA structural features on a genomic scale," vol. 41, pp. 56–62, 2013.

[32] S. E. Hunt *et al.*, "Ensembl variation resources," *Database*, vol. 2018, pp. 1–12, 2018.

[33] W. J. Kent *et al.*, "The Human Genome Browser at UCSC," vol. 12, pp. 996–1006, 2002.

[34] J. Casper *et al.*, "The UCSC Genome Browser database : 2018 update," vol. 46, no. November 2017, pp. 762–769, 2018.
[35] A. Siepel and D. Haussler, "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood," *Mol. Biol. Evol.*, vol. 21, no. 3, pp. 468–488, 2004.
[36] H. Zhao, Z. Sun, J. Wang, H. Huang, J. Kocher, and L. Wang, "CrossMap : a versatile tool for coordinate conversion between genome assemblies," vol. 30, no. 7, pp. 1006–1007, 2014.
[37] Turi, "Graphlab create." [Online]. Available: https://turi.com/index.html. [Accessed: 14-Mar-2017].
[38] E. Jones, T. Oliphant, and P. Peterson, "Scipy: Open Source Scientific Tools for Python." [Online]. Available: http://www.scipy.org. [Accessed: 03-Jun-2019].
[39] S. Seabold and J. Perktold, "Statsmodels : Econometric and Statistical Modeling with Python," in *9th Python in Science Conference*, 2010, pp. 57–61.
[40] T. C. Freeman *et al.*, "A gene expression atlas of the domestic pig," *BMC Biol.*, vol. 10, no. 90, 2012.
[41] R. A. Irizarry *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, 2003.
[42] T. Wang *et al.*, "Real-time resolution of point mutations that cause phenovariance in mice," *Proc. Natl. Acad. Sci. U. S. A.*, pp. E440–E449, 2015.
[43] M. J. Landrum *et al.*, "ClinVar : public archive of interpretations of clinically relevant variants," vol. 44, pp. 862–868, 2016.
[44] A. Ferna, J. M. Folch, L. Varona, R. Benı, and C. Rodrı, "Hypothalamic expression of porcine leptin receptor (LEPR), neuropeptide Y (NPY), and cocaine- and amphetamine-regulated transcript (CART) genes is influenced by LEPR genotype," *Mamm. Genome*, vol. 21, pp. 583–591, 2010.
[45] L. Fontanesi, A. Ribani, E. Scotti, V. J. Utzeri, N. Veličković, and S. D. Olio, "Differentiation of meat from European wild boars and domestic pigs using polymorphisms in the MC1R and NR6A1 genes," *Meat Sci.*, vol. 98, pp. 781–784, 2014.
[46] P. Latorre, C. Burgos, J. Hidalgo, L. Varona, J. A. Carrodeguas, and P. López-Buesa, "changes the enzyme kinetic and functional properties modifying fat distribution in pigs," *Sci. Rep.*, vol. 6, no. 19617, pp. 1–12, 2016.
[47] J. Ren *et al.*, "A missense mutation in PPARD causes a major QTL effect on ear size in pigs," *PLoS Genet.*, vol. 7, no. 5, 2011.
[48] M. F. L. Derks *et al.*, "Loss of function mutations in essential genes cause embryonic lethality in pigs," *PLoS Genet.*, vol. 15, no. 3, pp. 1–22, 2019.
[49] M. Chorev, A. Joseph Bekker, J. Goldberger, and L. Carmel, "Identification of introns harboring functional sequence elements through positional conservation," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, 2017.
[50] "Ensembl gene annotation update (e!90)," 2017. [Online]. Available: https://m.ensembl.org/info/genome/genebuild/2017_08_sus_scrofa_genebuild.pdf.
[51] "NCBI Sus scrofa Annotation Release 106," 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Sus_scrofa/106/. [Accessed: 29-Oct-2018].
[52] R. Buels *et al.*, "JBrowse : a dynamic web platform for genome visualization and analysis," *Genome Biol.*, vol. 17, no. 66, pp. 1–12, 2016.
[53] H. Hezroni, D. Koppstein, M. G. Schwartz, A. Avrutin, D. P. Bartel, and I. Ulitsky, "Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 Species," *Cell Rep.*, vol. 11, no. 7, pp. 1110–1122, 2015.
[54] R. Weikard, W. Demasius, and C. Kuehn, "Mining long noncoding RNA in livestock," *Anim. Genet.*, vol. 48, no. 1, pp. 3–18, 2017.
[55] K. de P. Lopes, F. J. Campos-Laborie, R. A. Vialle, J. M. Ortega, and J. De Las Rivas, "Evolutionary hallmarks of the human proteome: Chasing the age and coregulation of protein-coding genes," *BMC Genomics*, vol. 17, no. Suppl 8, pp. 337–349, 2016.
[56] A. B. Butler, "Evolution of vertebrate brains introduction and overview," *Encycl. Neurosci.*, vol. 4, pp. 57–66, 2009.
[57] K. Guschanski, M. Warnefors, and H. Kaessmann, "The evolution of duplicate gene expression in mammalian organs," *Genome Res.*, vol. 27, no. 9, pp. 1461–1474, 2017.

3.8. Appendix – Supplementary Data

Supplementary data files available online:

https://gsejournal.biomedcentral.com/articles/10.1186/s12711-020-0528-9

3.8.1. Annotation pre-processing

To train the pCADD model, SNVs from the generated training set were annotated with features assembled from various genomic annotations. The set of putative benign SNVs (derived alleles) represent mutations that are directed back in time while simulated variants are orientated forward in time. Therefore, annotations that are sensitive to these differences have to be swapped in the set of derived variants. Namely, the nucleotide reference and alternative columns (*Ref, Alt*), the amino acid substitutions (*nAA, oAA*) and the variant effect consequence predictions made by the ENSEMBL Variant Effect Predictor v91.3 for the labels *STOP Gained* and *STOP Lost*).

Not all SNVs were be able to be annotated with all genomic annotations, therefore missing values were imputed either by fixed values (such as 0.5, 1.0 or 0, False, UD) or by the mean of the SNVs in the simulated set. False was used for boolean values, UD (undefined) for factors. To deal with factors, all columns containing factor data were OneHotEncoded. This means factor data columns were replaced by as many columns with binary values as unique factors in these columns. In addition to the imputation, indicator columns were added to the data set which contain a 1 if a particular annotation is defined for a SNV or a 0 in the cases in which they do not. These genomic annotations for which indicator columns were created are: *cDNApos, CDSpos, protPos, SIFTval, Grantham* and *Dst2SplType_ACCEPTOR* & *Dst2SplType_DONOR*. The last two annotations are already OneHotEncoded data columns.

The annotations *minDistTSS* and *minDistTSE* were capped at 10000 and log transformed. The VEP consequences were summarized into 14 categories/factors (Table 2) and if there are multiple consequences per SNV, the category was chosen, following the order in Table 2.

Further, combinations of annotations were created. Namely, all possible combinations of *Ref* and *Alt* categories, generating an annotation for each possible nucleotide substitution. The same was done for *nAA* and *oAA*. Added to that, combinations of the 14 different VEP consequence summaries were formed with the following annotations: *cDNApos, CDSpos, Dst2Splice, GerpS, GerpN, lPhCons_noPig, mPhCons_noPig, verPhCons_noPig, lPhyloP_noPig, mPhyloP_noPig, verPhyloP_noPig, minDistTSS, minDistTSE, cDNApos, CDSpos, protPos, relcDNApos , relCDSpos , relprotPos , dnaHelT, dnaMGW, dnaProT, dnaProT*.

Before model training, all data columns were scaled by dividing each value by their column standard deviation.

3.8.2. Supplementary Tables

Table S1: Overview of genomic annotations which build the basis for features used to train the pCADD model. Overview and short description of genomic annotations and their imputed values in the case of missing data.

Annotation label	Data type	Imputed value	Annotation description
Ref	factor		Reference allele
Alt	factor		Observed allele
isTv	bool	0.5	Is transversion?
Consequence	factor		VEP Consequence summaries
GC	num	0.414	Percent GC in a window of +/- 75bp
CpG	num	0.023	Percent CpG in a window of +/- 75bp
motifECount	int	0.0	Total number of overlapping motifs
motifEHIPos	bool	False	Is the position considered highly informative for an overlapping motif by VEP
motifEScoreChng	num	0.0	VEP score change for the overlapping motif site
Domain	factor	UD	Domain annotation inferred from VEP annotation (ncoils, tmhmm, sigp, lcompl, ndomain = "other named domain")
Dst2Splice	int	0.0	Distance to splice site in 20bp; positive: exonic, negative: intronic
Dst2SplType	factor	UD	Closest splice site is ACCEPTOR or DONOR
oAA	factor	UD	Amino acid of observed variant
nAA	factor	UD	Reference amino acid
Grantham	int	0.0	Grantham score: oAA,nAA
SIFTcat	factor	UD	SIFT category of change
SIFTval	num	0.0	SIFT score
cDNApos	int	0.0	Base position from transcription start
relcDNApos	num	0.0	Relative position in transcript
CDSpos	int	0.0	Base position from coding start
relCDSpos	num	0.0	Relative position in coding sequence
protPos	int	0.0	Amino acid position from coding start
relProtPos	num	0.0	Relative position in protein codon
dnaRoll	num	0.255	Predicted local DNA structure effect on dnaRoll
dnaProT	num	0.518	Predicted local DNA structure effect on dnaProT
dnaMGW	num	0.0365	Predicted local DNA structure effect on dnaMGW
dnaHelT	num	-0.102	Predicted local DNA structure effect on dnaHelT
GerpS	num	-0.805	Rejected Substitution' score defined by GERP++
GerpN	num	1.384	Neutral evolution score defined by GERP++
GerpRS	num	0.0	Gerp element score
GerpRSpval	num	1.0	Gerp element p-Value
lPhCons_noPig	num	0.143	6-taxa-Laurasiatheria PhastCons score (excl. pig)
mPhCons_noPig	num	0.135	25-taxa-Mammalian PhastCons score (excl. pig)
verPhCons_noPig	num	0.126	100-taxa-Vertebrate PhastCons score (excl. pig)
lPhyloP_noPig	num	0.078	6-taxa-Laurasiatheria PhyloP score (excl. pig)
mPhyloP_noPig	num	0.106	25-taxa-Mammalian PhyloP score (excl. pig)
verPhyloP_noPig	num	0.294	100-taxa-Vertebrate PhyloP score (excl. pig)
minDistTSS	int	10000000	Distance to closest Transcribed Sequence Start (TSS)
minDistTSE	int	10000000	Distance to closest Transcribed Sequence End (TSE)

Table S2: Overview of the pig populations used in this study. List of pigs whose high frequency SNPs were added to the set of the putative benign (derived) variants to generate the training set. SNPs were called based on whole genome sequence data.

Number of individuals	Race/Breed
2	Angler Sattelschwein
2	Berkshire
2	British Saddleback
2	Bunte Bentheimer
1	Calabrese
7	Cassertana
2	Chato Murciano
8	Chinese Wild boar
2	Cinta Senese
53	Duroc
2	Gloucester Old Spot
10	Hampshire
11	Japanese Wild boar
3	Jiangquhai
2	Jinhua
43	Landrace
2	Large Black
97	Large White
2	Leping_spotted
2	Linderodsvinn
7	Mangalica
10	Meishan
2	Middle White
3	Negro Iberico
1	Nera Siciliana
13	Pietrain
3	Retinto
38	Synthetic
2	Tamworth
2	Thai domesticated pig
2	Thai Wild boar
2	Wannan spotted
37	European Wild boar
2	Xiang pig
1	Zang pig
4	NA

Table S3: VEP consequences summaries. VEP consequences summarized to 14 categories. If multiple annotations exist for the same variant, the consequence is selected according to the displayed hierarchy, starting at 1 and ending at 14.

Hierarchy	Abbreviation	VEP Consequence Summary
1	SG	Stop Gained
2	CS	Canonical Splice
3	NS	Non-Synonymous
4	SN	Synonymous
5	SL	STOP Lost
6	S	Splice Site
7	U5	5'-UTR
8	U3	3'-UTR
9	IG	Intergenic
10	NC	Noncoding-change
11	I	Intronic
12	UP	Upstream
13	DN	Downstream
14	O	Unknown

Table S4: Conservation score coverage of the pig genome. Coverage of the pig genome for the conservation scores used in the pCADD model (Supplementary Table 1). Y-chromosome, mitochondrial and unplaced scaffolds were excluded in pCADD and the conservation score calculations.

Conservation score	Nr. of positions	Fraction of the total genome
6-taxa-Laurasiatheria PhyloP score (excl. pig)	1,777,718,741	0.71
25-taxa-Mammalian PhyloP score (excl. pig)	1,978,673,774	0.79
100-taxa-Vertebrate PhyloP score (excl. pig)	1,367,857,535	0.55
GERP	1,043,440,638	0.42
6-taxa-Laurasiatheria PhastCons score (excl. pig)	1,777,718,741	0.71
25-taxa-Mammalian PhastCons score (excl. pig)	1,978,669,505	0.79
100-taxa-Vertebrate PhastCons score (excl. pig)	1,390,499,379	0.56
Golden Path Sscrofa11.1	2,501,912,388	

3.8.3. Supplementary figures

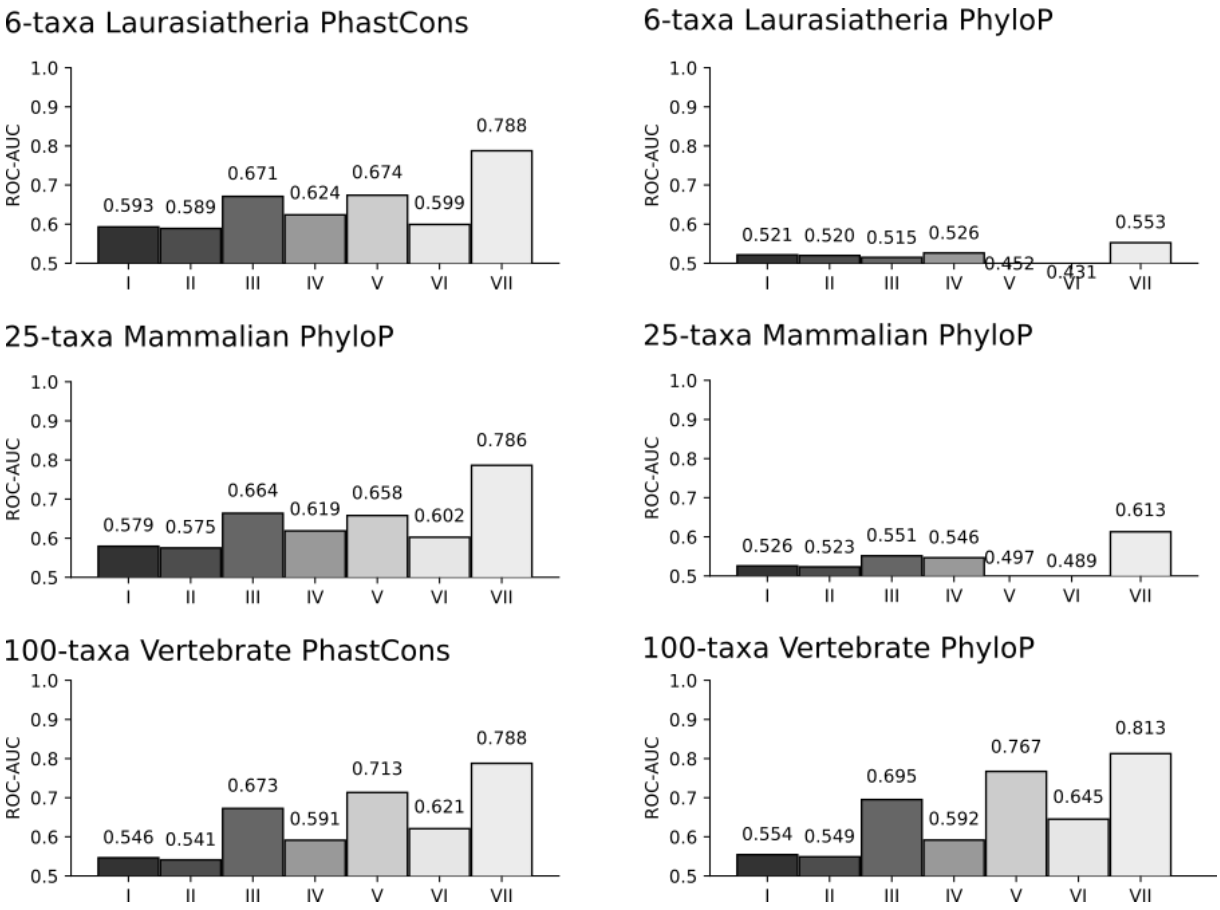


Figure S1: Prediction performances of six conservation scores on test sets, representing different regions of the genome for which different number of features are available. I: Whole test set; II: Intergenic SNVs; III: Transcribed SNVs; IV: SNVs in intron, 5' & 3' UTRs; V: Coding SNVs; VI: SNVs causing synonymous mutations; VII SNVs causing missense mutations.

Figure S2: Codon redundancy displayed in the JBrowser genome browser using pCADD scores.The third position in a codon is more redundant than either of the other two positions. This is reflected in the scores, here an example of the end of the 2nd exon of MACC1. MACC1 is located on the reverse strand.

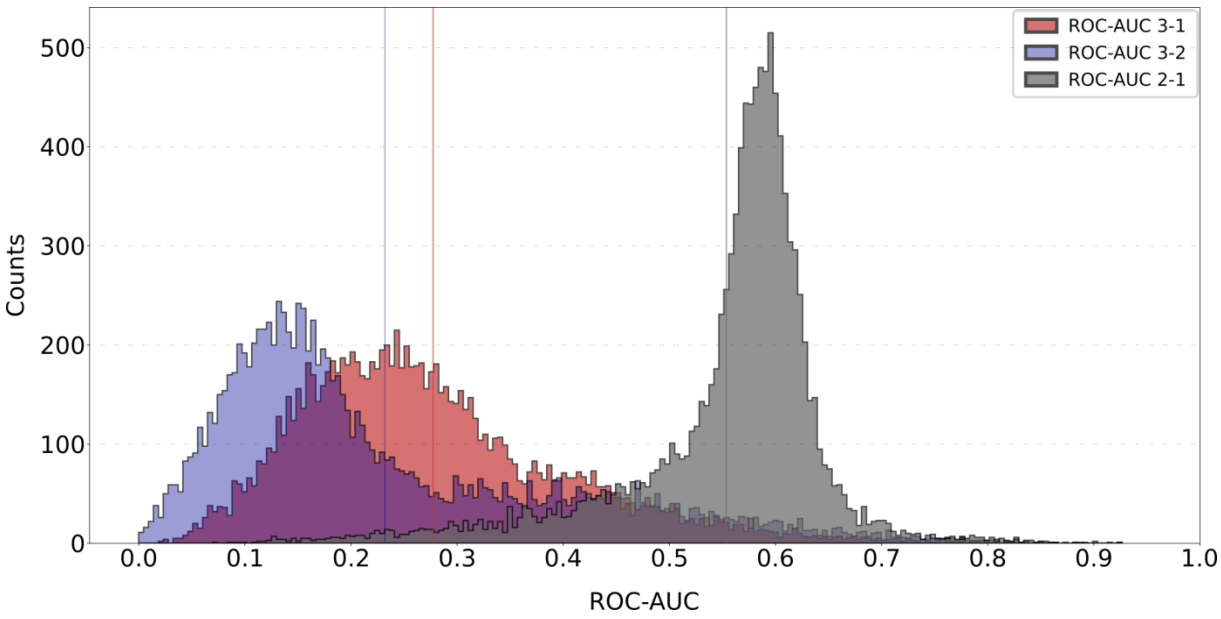


Figure S3: Effect sizes measured as ROC-AUC between the difference of pCADD scores of the three codon sites for all transcripts. The pCADD scores for the third and second codon positions differ generally the most (mean of ~0.232), thus their effect sizes have the largest absolute distance to 0.5. A ROC-AUC of 0.5 would indicate that no set of scores is larger than the other. The score indicates that the third position has a generally lower pCADD scores than the second position. The effect sizes of pCADD scores between the third and first codon positions (mean ROC-AUC ~0.277) also indicate that the third position is generally evaluated to be less deleterious than the first. In contrast, effect sizes between the second and first codon position are on average larger than 0.5 (mean of ~0.554) with the second codon position having a generally higher pCADD score than the first, which confirms that the second codon position is the most consequential when mutated. The effect sizes between the third and second codon positions as well as the third and first codon positions are more dispersed than between the second and first codon positions, probably due to the relatively larger variance in impact of a change at the third position than at the other two positions.

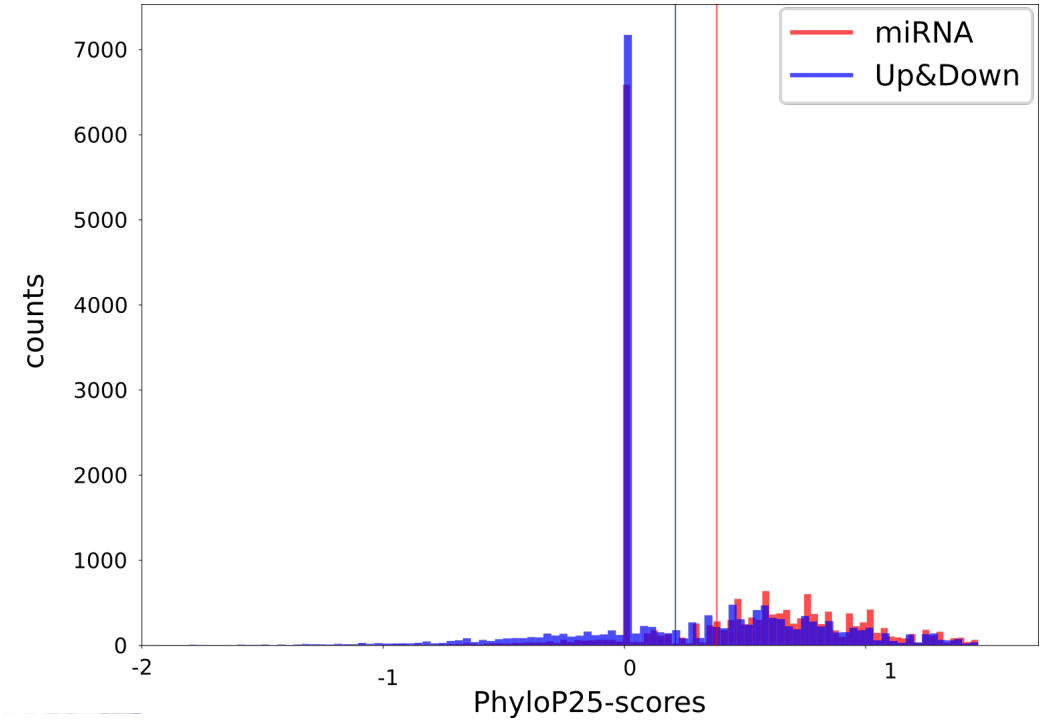


Figure S4: Histogram of conservation score distribution of (pre-)miRNA transcripts and their surrounding up- and downstream regions. Vertical lines indicate the mean values of each distribution with a mean of 0.382 for miRNA and 0.211 for Up&Down. The one-tailed Mann-Whitney U-test between both distributions returned a p-value of 1e-225 and a CLES of 59.54%. The conservation score used to annotate the transcripts and their surrounding regions are the 25-taxa-Mammalian PhyloP score shown in Supplementary Table 4.

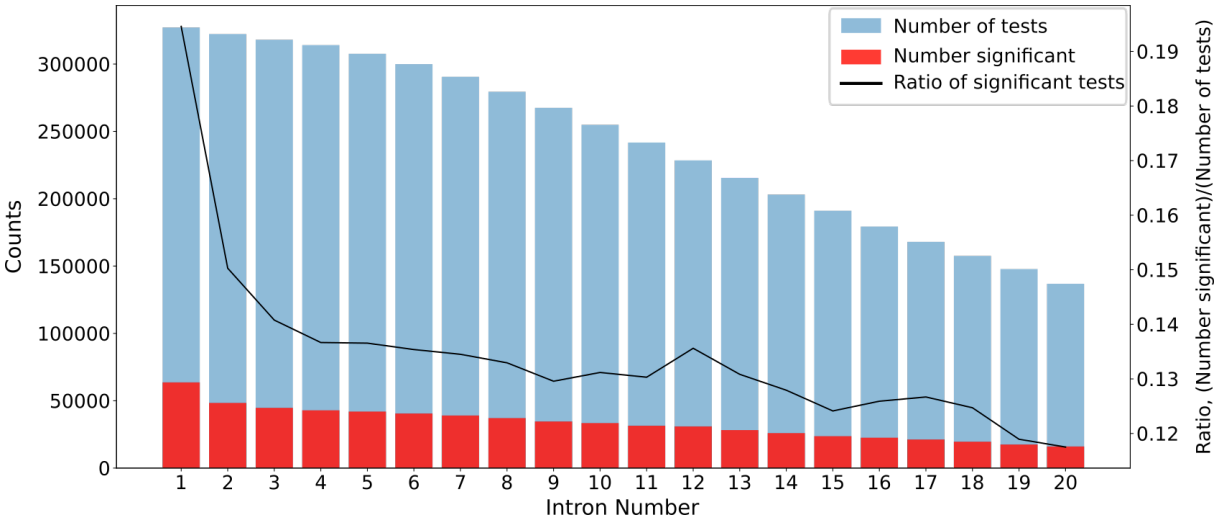


Figure S5: Comparison of the 25-taxa-Mammalian PhyloP scores per intron compared to all other introns, for the first 20 introns. The blue bar indicates the number of introns tested against the intron of interest, the red bar how many of these tests resulted in an adjusted p-value < 0.05 (scale on the left axis). As the intron position increases, the number of tests that can be conducted decreases (with the number of transcripts that have at least that many introns). In black, the normalised number of significantly enriched introns, normalized by the number of conducted tests per intron position (scale on the right axis).

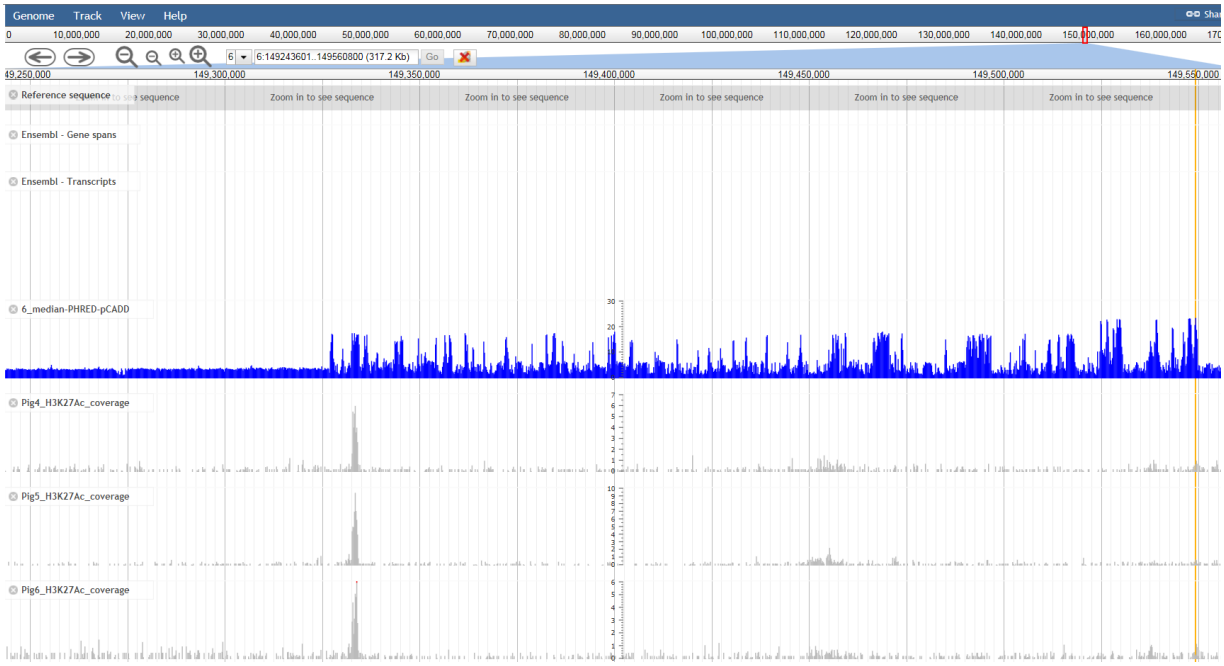


Figure S6: pCADD scores show a pattern of high scores in a presumably intergenic region. The yellow bar is indicating the location of the SNV 6:149549021T>C. It is embedded in a presumably intergenic region without any gene annotations in the pig genebuild of Ensembl and NCBI and the Ensembl genebuild of human when mapped to the human genome. The region is spiked with islands of high pPHRED scores, untypical for intergenic regions, and starts with an active enhancer region (peaks in H3K27Ac, data not part of this manuscript). The region 5' of the enhancer site is displaying patterns as expected for intergenic regions.

4. Accelerated discovery of functional genomic variation in pigs

Martijn F.L. Derks*

Christian Groß*

Marcos S. Lopes

Marcel Reinders

Mirte Bosse

Arne B. Gjuvsland

Dick de Ridder

Hendrik-Jan Megens

Martien A.M. Groenen

*shared first authorship

This chapter has been submitted for publication

4.1. Abstract

The genotype-phenotype link is a major research topic in the life sciences but remains highly complex to disentangle. Part of the complexity arises from the polygenicity of phenotypes, in which many (interacting) genes contribute to the observed phenotype. Genome wide association studies have been instrumental to associate genomic markers to important phenotypes. However, despite the vast increase of molecular data (e.g. whole genome sequences), pinpointing the causal variant underlying a phenotype of interest is still a major challenge, especially due to high levels of linkage disequilibrium.

In this study, we present a method to prioritize genomic variation underlying traits of interest from genome wide association studies in pigs. First, we select all sequence variants associated with the trait. Subsequently, we prioritize variation by utilizing and integrating predicted variant impact scores, gene expression data, epigenetic marks for promotor and enhancer identification, and associated phenotypes in other (well-studied) mammalian species. The power of the approach heavily relies on variant impact scores, for which we used pCADD, a tool which can assign scores to any variant in the genome including those in non-coding regions. Using our methodology, we are able to substantially narrow down the list of potential causal candidates from any association result. We demonstrate the efficacy of the tool by reporting known and novel causal variants, of which many affect (non-coding) regulatory sequences associated with important phenotypes in pigs.

This study provides an approach to pinpoint likely causal variation and genes underlying important phenotypes in pigs, accelerating the discovery of new causal variants that could be directly implemented to improve selection. Finally, we report several pathways and molecular mechanisms affecting important phenotypes in pigs, that can be transferred to human phenotypes.

4.2. Background

Closing the gap between genotype and phenotype is a major goal in many life sciences, but remains extremely challenging [1]. Part of the complexity arises from the polygenicity of phenotypes, in which many (interacting) loci contribute to the observed phenotype. Genome wide association studies (GWAS) have been instrumental to associate genomic markers to important phenotypes reported as quantitative trait loci (QTL), and to get a better grip on the biology of the traits [2]. However, the resolution of GWAS is limited by the correlation between neighbouring markers in linkage disequilibrium (LD). Hence, unravelling the molecular drivers underlying phenotypes of interest requires the identification of the actual causal variants [3], which often reside in the noncoding regions of the genome, in particular in predicted transcriptional regulatory regions [4].

In human genetics, a combination of statistical fine-mapping methods and expression QTL (eQTL) studies are used to further narrow down the list of candidate causal variants [5]. Further functional annotation, facilitated by large consortium efforts like the Encyclopedia of DNA Elements(ENCODE) [6], is used to prioritize variants based on likelihood of affecting a regulatory region, affecting gene expression. Despite this effort, identifying the causal variant remains difficult, partly because of the fundamental complexity of phenotype-genotype relations, in which also the environment plays an important role.

Also, in livestock, economically important phenotypes are typically determined by a very large set of variants each explaining a small fraction of the phenotypic variation. However, for many traits there are also some QTLs explaining a larger fraction (>1%) of the variation. For such larger QTLs

it is of interest to identify the underlying causal variation. Due to intense selection, the effective population size (Ne) of most livestock populations is small [7]. This often leads to extended LD, comprising up to millions of base pairs (Mb) in length, especially in regions with low recombination rates [8]. High LD yields an additional layer of complexity to fine-map GWAS results in livestock populations, and the use of crossbreeding to break down the LD is a costly, labour-intensive and time-consuming procedure to fine map the QTL region. On the contrary, livestock populations are less confounded by population stratification (i.e. ancestry differences between cases and controls), which can be a major factor in human GWAS studies [9].

Similar to human, further functional genomic information could help to prioritize the variants underlying the phenotypes of interest in livestock [10]. However, in pigs, the level of functional genomics information is limited. Fortunately, recent advances have been achieved in pigs by the publication of the pig Combined Annotation-Dependent Depletion (pCADD) tool [11], providing impact scores of any nucleotide substitution in the pig genome. CADD was developed to score variants with respect to their putative deleteriousness to prioritize potentially causal variants in genetic studies [12]. This tool is frequently used to score variants in human GWAS studies [5]. Subsequently, other species-specific CADD tools were developed [13]. The tool scores the deleteriousness (or functional impact) of single nucleotide variants (SNPs) and is built on many layers of annotations including sequence context, conservation scores, gene expression data, non-synonymous mutation scores, and epigenomic data, if available for the investigated species. The pCADD scores are the -10log10 of the relative rank of the investigated SNP among all possible SNPs in the Sus scrofa reference genome, giving the predicted 90% least impactful SNPs a pCADD score between 0-10, the least 99% a score between 0-20, et cetera.

Pig populations have been under a long-term biological experiment by animal breeders that use genomic selection to constantly improve their stock [14]. In general, genomic selection uses a variant panel on a chip to associate regions in the genome with important traits. This variant panel is distributed across the genome and allows within-population genetic variation to be captured [15]. However, genomic selection uses the genome as a "black box", as the SNPs on the chip are mostly not causal, but genetically linked to the actual causal variants and genes [16]. Therefore, the efficacy of genomic selection can be substantially improved by adding new genetic markers comprising the actual causal variation [17], providing insight in the exact molecular drivers involved in the selection.

The objective of this study is to bridge the genotype-phenotype gap in pig populations by pinpointing causal variants that are selected by genomic selection. More specifically, we will demonstrate that pCADD scores can be used to identify causal variants underlying GWAS peaks and QTLs. Being able to identify causal variants will have major implications for genomic selection and provides insights into the molecular biology and pathways affecting important phenotypes in pigs, that can be transferred to human phenotypes.

4.3. Results

4.3.1. Genome wide association studies in four elite pig populations reveal many QTLs affecting production, reproduction, and health

We analysed large scale genotype and phenotype data in four purebred pig populations: two boar breeds of Duroc and Synthetic origin, and two sow breeds of Landrace and Large White origin. In pigs, selection takes place on the purebred populations, while the final production animals are derived from three-way crosses. First, crossbred sows are created from populations selected for

high reproductivity and mothering abilities, which are subsequently crossed with a population especially selected for meat production traits. The examined traits can be grouped in three classes: (1) traits focussing on carcass and meat quality, including backfat, intramuscular fat, and growth; (2) reproduction traits, mainly focussing on litter size, number of liveborn, survival, and mothering abilities; and (3) health and welfare traits including disease resistance, osteochondrosis, umbilical hernia, and other conformation traits. A total of 129,336 animals with 552,000 imputed SNPs were subjected to a GWAS analysis for 83 traits. The analysis revealed a large set of QTL regions with a genome-wide association significance threshold of -log10(p)>6.0, and significant associations were observed for the majority of examined traits. The 'lead' SNP that showed the strongest association signal is used as a starting point for further analysis.

4.3.2. A pipeline for integrating pCADD scores and functional information to rank sequence variants

4.3.2.1. pCADD evaluates all possible substitutions from the Sscrofa11.1 pig reference genome

Our approach first relies on the lead SNP from a significant GWAS peak to extract sequence variants that are in high LD (r2>0.7). The whole-genome sequence variants are extracted from a total of 428 animals (Duroc: 101, Synthetic: 71, Landrace: 167, Large White: 89), sequenced to an average depth of 11.82. Next, we assigned pCADD scores to each sequence variant in high-LD with the lead SNP, to prioritize them on their likely impact. The sequence variants were assigned to a functional class using the Ensembl Variant Effect Predictor (VEP, release 98) [18]. The distribution of the pCADD scores for a set of variants depends on their functional class, and non-coding variants have on average lower scores compared to coding variants. The quantiles and further class statistics for the pCADD scores are presented in Table S1. In addition, three liver histone modification datasets were used (for modifications H3K27Ac and H3K4me3) to mark variation overlapping with regulatory sequences, including likely active promoter and enhancer elements in pig liver tissue [19].

4.3.2.2. Phenotype and pathway information provides further evidence of gene causality

Functional annotations, including pathways and gene-ontology information for the examined pig genes associated with the top-ranked variants, were extracted from the Uniprot database [20]. Moreover, we extracted associated phenotypes from orthologous genes from the Ensembl database for human (Homo sapiens), mouse (Mus musculus), and rat (Rattus norvegicus). The phenotypes are mainly based on (disease) association studies in human, and gene-knockouts in mouse and rat [21]. A complete overview of the pipeline is presented in Figure 1.

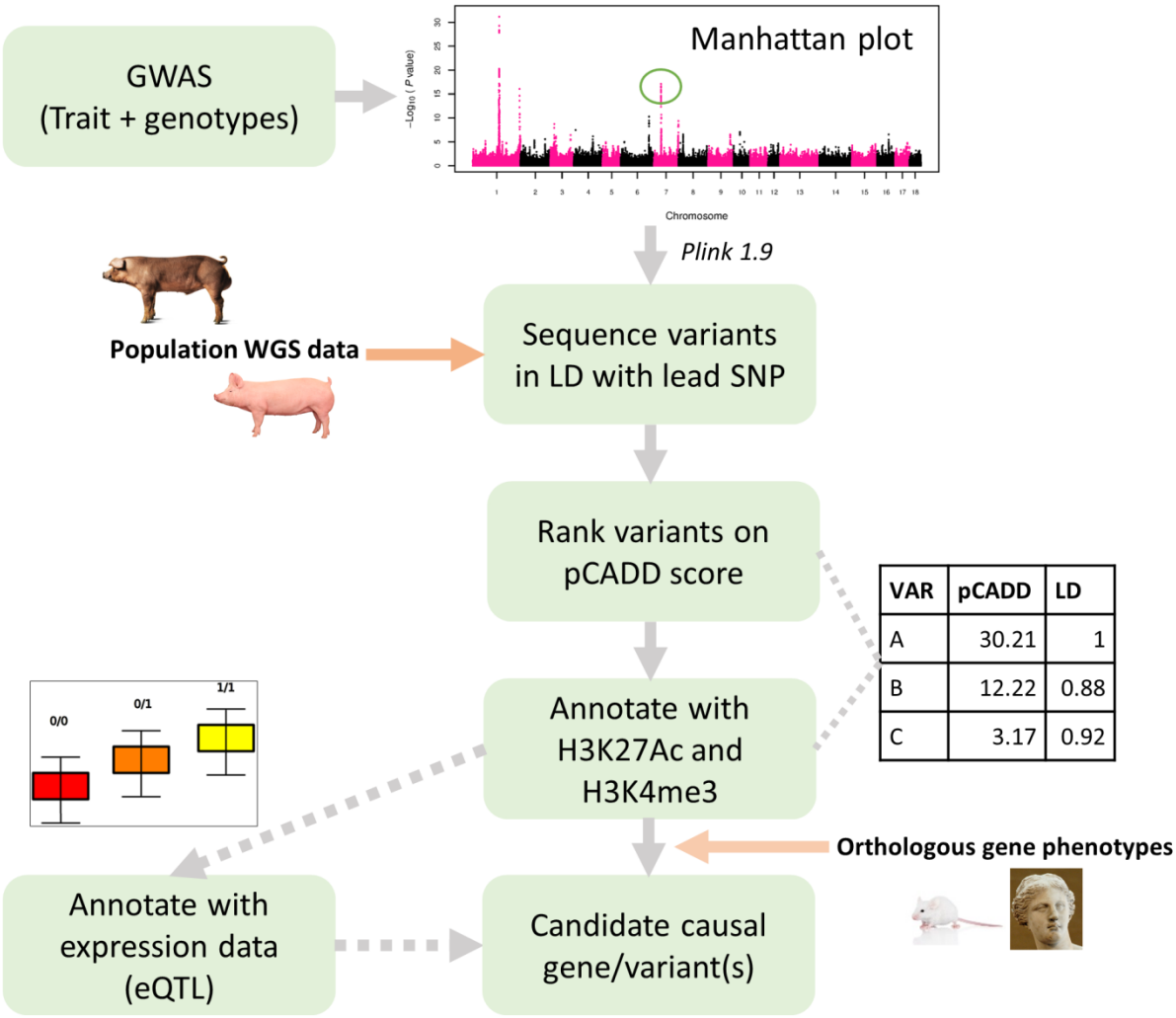


Figure 1: Pipeline overview. The pipeline takes the result of a GWAS as input (lead SNP) and identifies SNPs from WGS data that are in high LD with the lead SNP. Subsequently, the variants are prioritizes based on impact scores (pCADD), open chromatin information (liver), and gene expression (if available). The pipeline outputs a final list of candidate causal variants for each trait of interest, ranked on its likely importance.

4.3.2.3. Gene expression information allows identification of possible expression quantitative trait loci

The combination of genotype and gene expression data provides an additional layer of evidence to find causal variation, as differences in expression of genes can be associated with a variant (expression quantitative trait loci; eQTL). In this study we use 59 RNA-sequenced samples [22] from Landrace (n=34) and Duroc (n=25) to test for differential expression between the genotype classes (homozygous reference, heterozygous, homozygous alternative) to associate the expression of genes with the genotypes. The samples were sequenced from testis tissue, further details about the sequenced samples and alignment depth are provided in Table S2. The combination of epigenomic marks (liver) and gene-expression data (testis) can, on top of the pCADD scores, facilitate in the discovery of functional variants.

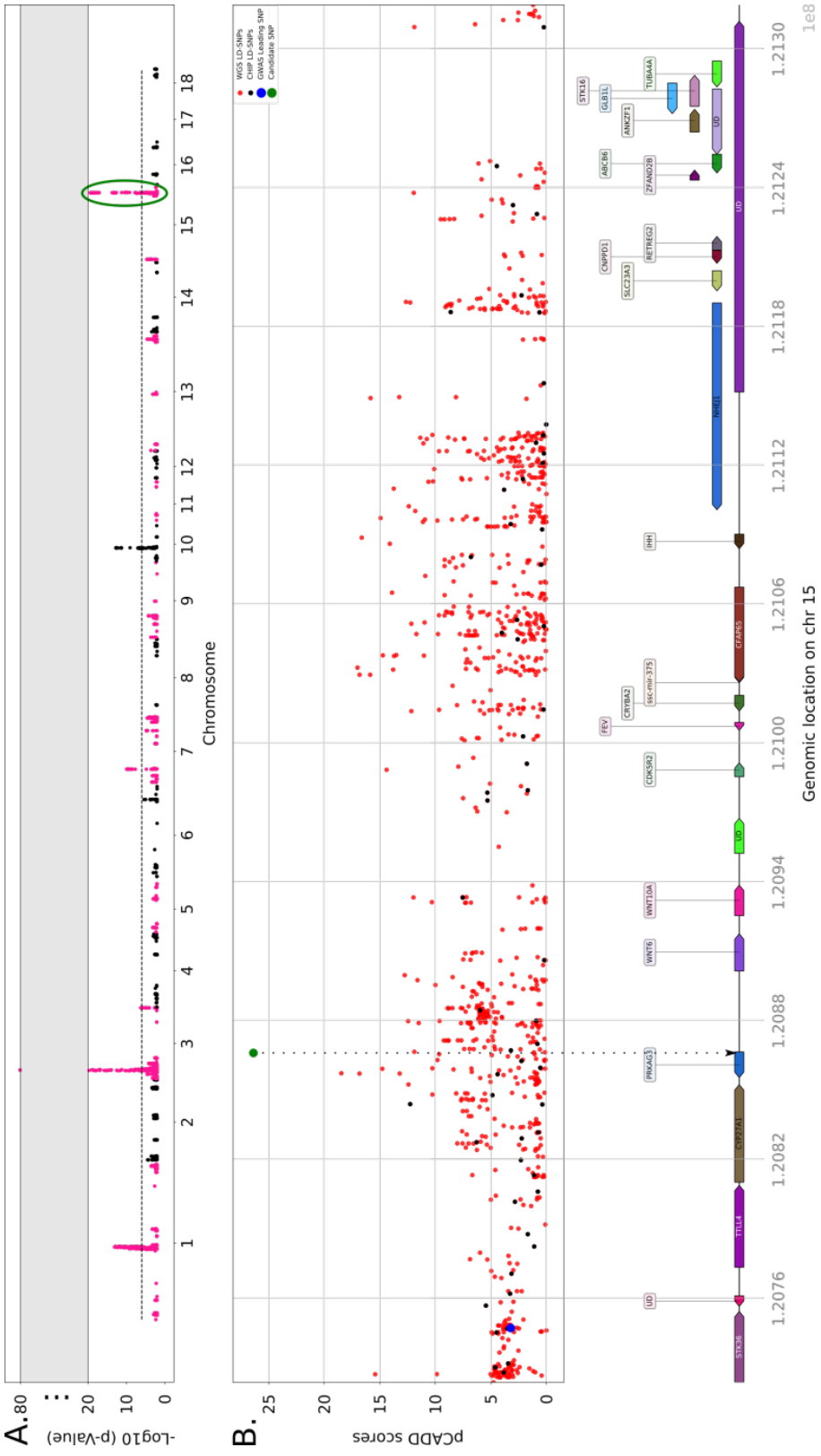


Figure 2: A) Manhattan plot for drip loss in Duroc showing a strong QTL on chromosome 15:121Mb. Only SNPs with a -log$_{10}$(p) > 2 are plotted. B) Plot showing all sequence variants in high LD (red) with the lead SNP (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98.

4.3.3. Accelerated discovery of potential causal variants from GWAS results

To demonstrate the utility of our approach we first analysed several QTL regions with known causal variants reported in literature. This list includes a missense mutation in MC4R affecting production traits [23], a promoter variant affecting number of teats in the VRTN gene [24], and a missense mutation affecting meat quality in PRKAG3 [25]. The method returned the causal variant as top ranked for both the MC4R missense mutation (Text S1, Figure S1) and the VRTN promoter variant (Text S2, Figure S2), despite the fact that hundreds of variants were found in LD with the lead SNP.

The mutation identified by Milan [25] does not segregate in our sequenced animals, however, we identified another missense variant (15:g.120865869C>T) in the PRKAG3 gene likely affecting meat quality in both boar breeds (Figure 2), as described by Uimari et al. 2014 [26]. The causal missense variant is highlighted in green, and the lead SNP in the GWAS results in blue in Figure 2B. The variant substitutes glutamic acid for lysine (ENSSSCP00000030896:p.Glu47Lys) and is segregating at a frequency of approximately 20%, and 36% in Synthetic and Duroc, respectively. PRKAG3 regulates several intracellular pathways, including glycogen storage [27]. The specific isoform (ENSSSCT00000036402.2) affected by the Glu47Lys missense mutation has a role in the metabolic plasticity of fast-glycolytic muscle and is primarily expressed in white skeletal muscle fibers [28]. Gain of function mutations in the PRKAG3 gene have been correlated with increased glycogen content in skeletal muscle in pig, negatively affecting meat quality [29]. The Lys47 variant likely causes a gain-of-function of the 5'-AMP-activated protein kinase subunit gamma-3 enzyme, resulting in increased glycogen content causing lower water holding capacity resulting in low meat quality.

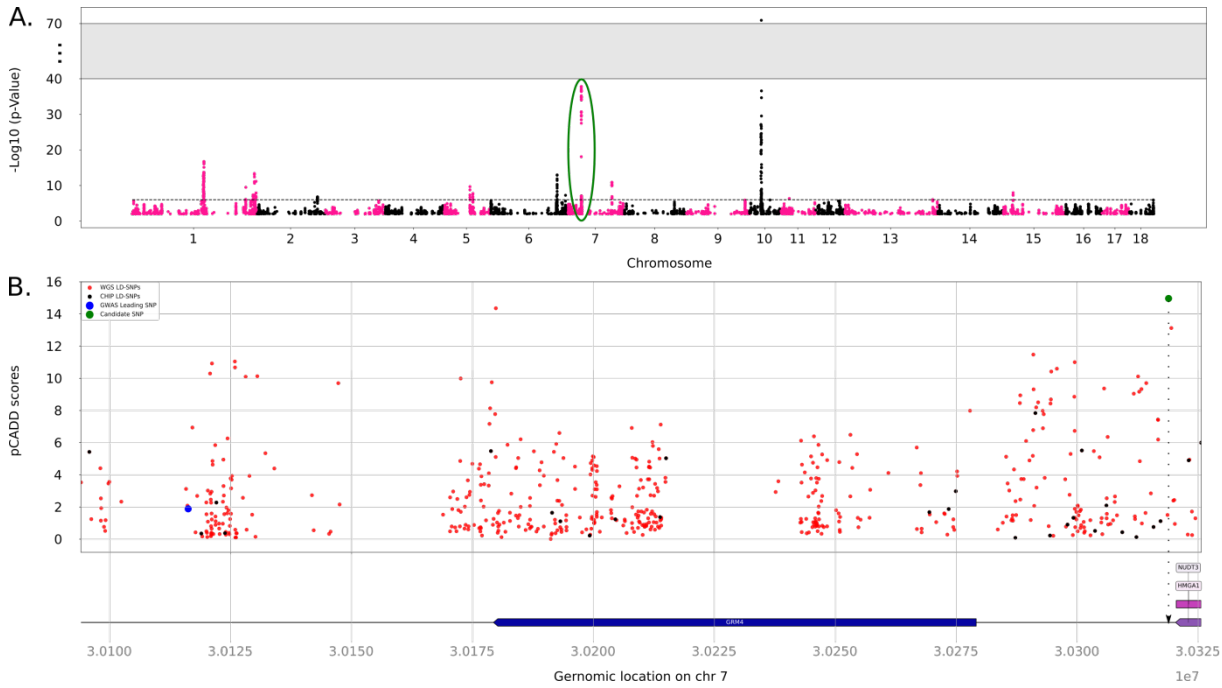


Figure 3: A) Manhattan plot for backfat in Duroc showing a strong QTL on chromosome 7:30Mb. Only SNPs with a -log$_{10}$(p) > 2 are plotted. B) Plot showing all sequence variants in high LD (red) with the lead SNP (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98.

4.3.4. Large scale analysis reveals several novel variants with pleiotropic effects on important phenotypes

4.3.4.1. Promoter variants in the HMGA1 and HMGA2 genes affect fat deposition and growth in pigs

A strong QTL on chromosome 7 affects backfat, intramuscular fat, growth, feed intake and loin depth in Duroc (Figure 3A). The lead SNP in the GWAS result is located at position 7:30,116,227 with a -log10(p) > 20 for backfat, feed consumption, and intramuscular fat (Figure S4). The analysis returns 485 variants in high LD with the lead SNP (Figure 3B). The two variants with the highest pCADD scores are annotated upstream of the HMGA1 gene, 566 bp apart (Figure 3B). Both mutations are in the promoter region of the HMGA1 gene, supported by signals on the H3K4me3 and H3K27Ac histone marks (Figure S5). The A allele, segregating at 36% allele frequency, is associated with less backfat, faster growth, but also smaller loin and decreased intramuscular fat. We evaluated the expression of the HMGA1 gene in twenty samples for which both genotype and gene expression, as normalized fragments per kilobase per million (FPKM), were available within the three genotype classes GG, AG, and AA. The A allele causes increased expression of the gene in an additive manner (P=0.041, Figure S6) and suggests that increased expression of the HMGA1 gene positively affects backfat and growth, but decreases intramuscular fat. In addition, we find two variants affecting the promoter region of the HMGA2 gene, to be associated with less backfat in the Synthetic breed (Table 1). Both HMGA1 and HMGA2, part of the High Mobility Group A gene family, are well-known genes to affect growth and stature in pigs [30]–[32], but no causal variant has been reported thus far. Our results suggest that the causal variants for both genes are regulatory.

Table 1: List of potential causal variants identified from the pipeline. Table shows the variants type, potential overlap with promotor or enhancer region (from liver [19]), the change in amino acid (for missense mutations) and the pCADD score for variants affecting one or more important selection traits (BFE: backfat, IMF: intramuscular fat, TGR: growth rate, DRY: drip loss, NTE: number of teats). The causal variant for genes in bold have already been reported in literature.

Chr	Variant	Type	Promotor/	Amino acid	pCADD score	Rank	Gene	Breed(s)	BFE	IMF	TGR	DRY	NTE	Supp.
1	G-263595807-T	missense	NO	A252D	20.05	2	*COPS4*	Duroc	+	NS	-	NS	NS	[34]
6	A-145977262-T	intron	NO	-	14.63	1	*SGIP1*	Large White	NS	NS	-	NS	NS	[58]
12	C-44684331-G	missense	NO	G131R	25.81	1	*SLC46A*	Synthetic	NS	NS	+	NS	NS	[34]
9	C-17077403-A	5' UTR	YES	-	8.45	7	*PRCP*	Synthetic	NS	NS	-	NS	NS	[57]
5	A-83681067-G	intron	YES	-	12.1	9	*NR1H4*	Synthetic	NS	NS	-	NS	NS	[56]
2	G-15310202-A	3' UTR	NO	-	21.38	1	*NR1H3*	Synthetic	NS	NS	-	NS	NS	[55]
7	A-113911774-G	intron	NO	-	9.72	1	*JARID2*	Duroc	NS	+	NS	NS	NS	[54]
6	A-146830209-G	intron	NO	-	11.88	1	*LEPR*	Duroc	NS	+	NS	NS	NS	[53]
15	C-117292901-A	missense	NO	G1693C	24.75	1	*ABCA12*	Large White	NS	+	NS	NS	NS	[52]
4	A-88412353-C	intron	NO	-	18.99	1	*NOS1AP*	Large White	NS	+	NS	NS	NS	[51]
2	C-41019232-T	upstream	NO	-	3.91	9	*SAA3*	Large White	NS	+	NS	NS	NS	[50]
2	T-103610859-C	missense	NO	L335S	21.45	1	*LNPEP*	Synthetic	NS	+	NS	NS	NS	[49]
14	G-128748846-A	5'UTR	YES	-	15.53	7	*CACUL1*	Synthetic	NS	-	NS	NS	NS	[48]
3	C-94863278-A	5'UTR	YES	-	16.11	7	*PRKCE*	Landrace	-	NS	NS	NS	NS	[47]
13	A-195332161-G	intron	YES	-	6.13	26	*SOD1*	Duroc	-	NS	NS	NS	NS	[46]
11	T-20619202-C	3'UTR	NO	-	18.46	1	*HTR2A*	Duroc	-	NS	NS	NS	NS	[45]
5	G-65814519-A	missense	NO	V850I	23.1	1	*AKAP3*	Duroc	+	NS	NS	NS	NS	[44]
2	A-144841051-C	intron	NO	-	10.65	2	*NR3C1*	Synthetic, Large [illegible]	+	+	NS	NS	NS	[43]
18	T-10098558-C	intron	NO	-	16.41	1	*HIPK2*	Synthetic	-	-	NS	NS	NS	[42]
8	A-102781174-G	missense	NO	M165V	21.27	1	*QRFPR*	Synthetic	NS	NS	NS	NS	-	[41]
7	A-97614602-C	upstream	NO	-	11.95	2	*VRTN*	Duroc, Landrace, [illegible]	NS	NS	NS	NS	+	[24]
15	A-46758359-G	intron	YES	-	11.73	4	*SORBS2*	Synthetic	NS	NS	NS	-	NS	-
14	T-107058908-C	intron	YES	-	24.5	1	*SORBS1*	Synthetic	NS	NS	NS	-	NS	[40]
9	G-758928-A	missense	NO	A773T	21.92	1	*TRIM66*	Synthetic	NS	NS	NS	-	NS	[39]
13	G-17363457 6-A	upstream	YES	-	15.68	1	*GBE1*	Synthetic	NS	NS	NS	+	NS	[38]
9	C-9329652-T	missense	NO	P419S	20.91	3	*NEU3*	Synthetic	NS	NS	NS	-	NS	[37]
2	C-96202720-T	intron	NO	-	17.86	2	*MEF2C*	Synthetic	NS	NS	NS	-	NS	[36]
1	C-127921686-T	missense	NO	G1904S	23.03	1	*MAP1A*	Synthetic	NS	NS	NS	-	NS	[35]
6	C-67433001-T	intron	YES	-	11.87	2	*KLHL21*	Landrace	NS	NS	NS	+	NS	[34]
15	C-120865869-T	missense	NO	E47K	26.37	1	*PRKAG3*	Synthetic, Duroc	NS	NS	NS	-	NS	[26]
5	T-30187091-C	upstream	NO	-	19.44	2	*HMGA2*	Synthetic	-	NS	NS	NS	NS	[32]
7	G-30318881-A	upstream	YES	-	14.96	1	*HMGA1*	Duroc	-	-	+	+	NS	[30]
1	G-160773437-A	missense	NO	D298N	27.47	1	*MC4R*	Synthetic, Duroc	-	NS	+	NS	[23]	
1	G-120074006-A	missense	NO	T386M	30.27	1	*SCG3*	Synthetic	-	+	NS	-	NS	[33]

Figure 4: A) Manhattan plot for backfat in the Synthetic breed showing a strong QTL on chromosome 1:116Mb. Only SNPs with a -log$_{10}$(p) > 2 are plotted. B) Plot showing all sequence variants in high LD (red) with the lead SNP (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98.

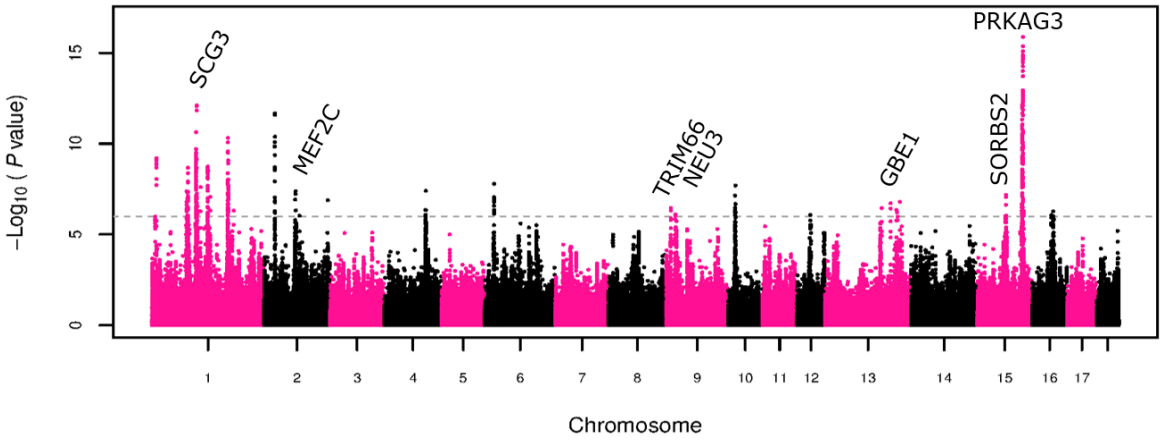


Figure 5: Manhattan plot for drip loss in the Synthetic breed. The figure shows significant loci and likely causal genes identified.

4.3.4.2. A novel missense mutation in SCG3 likely to affect backfat and growth rate

A strong QTL on chromosome 1 affects backfat, intramuscular fat, and drip loss in the Synthetic breed (Figure 4A). The lead SNP in the GWAS result is located at position 1:115,884,118. The analysis returns 874 variants in high LD with the lead SNP. The SNP with the highest pCADD score (1:g.120074006G>A), a single missense variant affecting the SCG3 gene is identified as the likely culprit (Figure 4B). The variant substitutes a threonine for a methionine at position 386 in the Secretogranin-III protein (ENSSSCP00000044507:p.Met386Thr). The Met386 allele is associated with increased intramuscular fat, more backfat and lower meat quality. Several variants affecting the SCG3 gene have been associated with obesity in humans [33], supporting its likely causality for the fat-associated phenotypes in pigs.

4.3.4.3. A novel missense mutation in COPS4 likely to affect backfat and growth rate

A QTL on chromosome 1 affects growth and backfat in the Duroc breed (Table 1). The lead SNP in the GWAS result is located at position 1:265,017,724. The analysis returns 706 variants in high LD with the lead SNP. The second pCADD-ranked SNP (1:g.263595807G>T), a single missense variant affecting the COPS4 gene is identified as likely causal. The variant substitutes an alanine for an aspartic acid at position 252 in the COP9 signalosome complex subunit 4 protein (ENSSSCP00000056478:p.Ala252Asp). The Asp252 allele is associated with less backfat and slower growth. Variants affecting COPS4 have been associated with increased body weight in mice [34].

4.3.4.4. Balancing selection for causal variants in the breeding program

Several identified variants exhibit pleiotropic effects for important selection traits, e.g. variants affecting HMGA1, SCG3, COPS4, and MC4R (Table 1). Variants that positively affect backfat often have negative consequences for growth, while variants that positively affect intramuscular fat often show detrimental effects on meat quality. The observed pleiotropic effects cause the variants to be under balancing selection in the breeding program, preventing population fixation of individual variants underlying strong QTL regions.

4.3.5. Variants affecting production and meat quality traits enriched for specific molecular mechanisms

4.3.5.1. Genes affecting meat quality involved in muscle glycogen storage

We identified several candidate causal variants that affect meat quality. Especially in the Synthetic breed, we find 26 loci significantly associated with drip loss (-log10(p) > 6), a meat quality trait that measures the water holding capacity of the meat (Figure 5). The top ranked pCADD-scored

genes show a strong enrichment for pathways involved in glycogen synthesis and storage (Table 1). Increased levels of muscle glycogen lead to increased drip loss, negatively affecting meat quality [59]. Examples of such variants include two regulatory variants affecting the MEF2C and GBE1 genes. MEF2C knockout mice accumulate glycogen in their muscles [36], while GBE1 codes for a glycogen branching enzyme associated with glycogen storage disease, if mutated [38]. Moreover, we identify two missense variants affecting the NEU3 (ENSSSCP00000034065:p.Pro419Ser) and MAP1A (ENSSSCP00000005070:p.Gly1904Ser) genes, both directly involved in the glycogen deposition [35], [37].

4.3.5.2. Genes affecting growth and fat deposition traits are involved in energy metabolism and adipogenesis

We identified several likely causal variants and genes affecting other important production traits (Table 1). The top-ranked genes are enriched in energy reserve metabolic processes, glycogen metabolic process, regulation of lipid biosynthetic process, and homeostasis (Table S3). More specifically, two identified regulatory variants in the SOD1 and PRKCE genes likely affect backfat. SOD1 is involved in glucose metabolism and prevents oxidative damage associated with obesity [46], while mutations in PRKCE decrease the amount of body fat [47]. Furthermore, we identified one regulatory variant in the CACUL1 gene affecting intramuscular fat. This gene inhibits adipogenesis via the peroxisome proliferator-activated receptor γ (PPARγ) [48]. In addition, two missense variants affect intramuscular fat via the LNPEP (ENSSSCP00000051249:p.Leu334Ser) and ABCA12 (ENSSSCP00000058038:p.Gly1693Cys) genes. LNPEP attenuates diet-induced obesity in mice through increased energy expenditure, and decreases the amount of adipose tissue [49], while the ABCA12 gene plays an important role in lipid transport, affecting carcass fat content in pigs [52]. We further identified regulatory variants in the NR1H3, NR1H4, and PRCP genes, all likely affecting growth (Table 1). NR1H3 and NR1H4 are paralogous genes both involved in lipid homeostasis [55], [56], while reduced levels of PRCP expression promote obesity by regulating the α-melanocyte-stimulating hormone (α-MSH) that regulates feeding behaviour. Finally, we found a missense variant in the SLC46A1 gene associated with increased intramuscular fat (ENSSSCP00000020843:Gly131Arg) in pigs, known to affect glucose and fat levels in knockout mice [34].

4.4. Discussion

The aim of this study was to prioritize variants associated with important traits in pigs. The variants are ranked based on pCADD scores, and possibly further supported with respect to their function by epigenetic marks and gene expression data. The method is especially relevant because genomic variation underlying phenotypic variation mostly affects the non-coding part of the genome [4], and GWAS results often point to regions outside gene boundaries (Bartonicek et al. 2017). With the publication of the pCADD scores [11], a powerful resource is now available to rank any possible substitution variant in the genome based on the likelihood of being functional. This is a major step forward in livestock, as thus far only variation in the coding region could be scored. On top of the pCADD scores, we use epigenomics and gene expression data to annotate regulatory sequences and associate gene expression to the trait of interest. In human, many transcriptomic and epigenomic marks have already been incorporated in the CADD scores [12]. However, the pCADD scores are built on far less (epi)-genomics data, but with the accumulation of functional genomic data in pigs [60], these pCADD scores will further improve.

Livestock populations generally have small effective population sizes (Ne: 50-200), far less compared to e.g. human (Ne ~ 10,000), leaving much longer blocks of variants in high LD. This high level of LD increases the power to detect QTL regions, even with relatively low SNP density. However, within large LD blocks, many variants will be associated, and a thorough variant

prioritization should be performed to point to likely causal variants within the (often) large variant set. For example, the LD block for the number of teats in Landrace spans about 1.8 Mb, leaving many thousands of variants in linkage, which increases the level of noise and hampers the detection of the causal variant. Nevertheless, in Large White and Duroc, which have smaller LD-blocks (100-500 kb), the causal VRTN promoter SNP is among the top SNPs. In that sense, integrating the results from multiple breeds provides additional power to further narrow down the list of candidates, assuming that the same causal variant is segregating, but likely with a very different underlying haplotype structure. This example shows that the tool can be very powerful to prioritize variants, but with a trade-off for the level of LD, increasing the noise when many thousands of variants are in linkage.

Although the development of genomic selection has revolutionized the world of animal breeding, the lack of functional genomic information currently limits further development [61]. The framework and associated pCADD scores provided within this study will accelerate the discovery of new functional variants, which can be directly implemented in genomic selection by adding the causal variants to the selection chip used for genomics selection. Moreover, the results provide further knowledge of the biological pathways associated with important phenotypic variation in livestock. For this, the (functional) genome annotation in livestock genomes is still of too low quality compared to other well-studied mammalian species [60]. Therefore, using annotations from human, mouse, and rat will often provide more detailed information on gene function, pathways, and associated genes compared to the pig annotation itself.

The populations under study provide an interesting framework to study common pathways and molecular mechanisms involved in comparable phenotypes between pig and human. For example, we report the GBE1 gene affecting meat quality in pigs by accumulating glycogen in the muscle, a gene associated with glycogen storage disease in human [62]. Moreover, several of the identified genes affecting growth and fat deposition traits in pigs are involved in energy metabolism, glucose homeostasis, and adipogenesis, often associated with metabolic disease in human (e.g. HMGA1, SCG3 genes). In human, however, environmental factors play a very large role in the formation of metabolic disease, while in pigs the animals are kept under relatively stable conditions, which could make the pig an ideal model to study the effects of specific genic variants on these analogous phenotypes [63]. Pig breeding has led to extreme changes in animal production and efficiency, with very little negative consequences on health [14]. This remarkable robustness of the animals, and the molecular mechanisms involved, could help to understand metabolic disease in human. Finally, our study implicates that, despite the complexity of pathways, there are several key entry points (i.e. genes) with a large effect on specific phenotypes in pigs, likely to be similar in human. Understanding these 'key' genes, and how they function together would further help to unravel the (molecular) consequences of genomic selection.

4.5. Conclusion

This study integrates pig CADD scores and various sources of functional data to provide a framework to pinpoint causal variation associated with important phenotypes in pigs. We demonstrate our method by identifying novel causal mutations or substantially narrow down the list of potential causal candidates in various strong QTL regions, affecting both production and reproduction traits. The new regulatory variants can be utilized directly in the breeding program to improve selection substantially, and to better understand the biology and molecular mechanisms underlying the selected traits. Finally, the pig populations under study provide an interesting framework to study common pathways and molecular mechanisms involved in analogous phenotypes between human and pig.

4.6. Methods

4.6.1. Ethics statement

Samples collected for DNA extraction were only used for routine diagnostic purpose of the breeding programs, and not specifically for the purpose of this project. Therefore, approval of an ethics committee was not mandatory. Sample collection and data recording were conducted strictly according to the Dutch law on animal protection and welfare (Gezondheids- en welzijnswet voor dieren).

4.6.2. Genotype data and breeds

The dataset consists of 15,791 (Duroc), 28,684 (Synthetic), 36,956 (Large White), and 41,865 (Landrace) animals genotyped on the (Illumina) Geneseek custom 50K SNP chip with 50,689 SNPs (50K) (Lincoln, NE, USA) and imputed to the Axiom porcine 660K array from Affymetrix (Affymetrix Inc., Santa Clara, CA, United States). The chromosomal positions were determined based on the Sscrofa11.1 reference assembly [64]. SNPs located on autosomal chromosomes were kept for further analysis. Next, we performed per-breed SNPs filtering using following requirements: each marker had a MAF greater than 0.01, a call rate greater than 0.85, and an animal call rate > 0.7. SNPs with a p-value below 1x10-12 for the Hardy-Weinberg equilibrium exact test were also discarded. All pre-processing steps were performed using Plink v1.90b3 [65].

4.6.3. Phenotypes

The phenotypes consisted of 1,360,453 records of purebred and crossbred offspring of genotyped animals from four lines of different origin: Duroc, Synthetic, Landrace, Large White.

4.6.4. Genome wide association study

A single SNP GWAS was performed with the software ASReml [66] by applying the following model:

$$\mathrm{DEBV_{ij}w} = \mu + \mathrm{SNP_i} + \mathrm{a_j} + \mathrm{e_{ij}}$$

where $DEBV_{ij}$ is the DEBV (deregressed estimated breeding value) of SNP *i* for genotyped animal *j*, μ is the overall DEBV mean of the genotyped animals, $\mathrm{SNP_i}$ is the genotype of the SNP *i* coded as 0, 1 or 2 copies of one of the alleles, a_j is the additive genetic effect and e_{ij} the residual error. The weighting factor *w* was used in the GWAS to account for differences in the amount of available information on offspring to estimate *DEBV* [67]. Association results were considered significant if -log10(p) > 6.0.

4.6.5. Population sequencing and mapping

Sequence data was available for 101 (Duroc), 71 (Synthetic), 167 (Landrace), and 89 (Large White) animals from paired-end 150 bp reads sequenced on Illumina HiSeq. The sequenced samples are frequently used boars, selected to capture as much as possible of the genetic variation present in the breeds. The coverage ranges from 6.6 to 22.2, with an average coverage of 11.82. FastQC was used to evaluate read quality [68]. BWA-MEM (version 0.7.15 [69]) was used to map the WGS data to the Sscrofa11.1 reference genome. SAMBLASTER was used to discard PCR duplicates [70], and samtools was used to merge, sort, and index BAM alignment files [71].

4.6.6. Variant discovery functional class annotation

FreeBayes was used to call variants with following settings: --min-base-quality 10 --min-alternate-fraction 0.2 --haplotype-length 0 --ploidy 2 --min-alternate-count 2 [72]. Post processing was performed using BCFtools [69]. Variants with low phred quality score (<20), low call rate (<0.7) and variants within 3 bp of an indel are discarded, leaving a total of 21,648,132 (Landrace), 23,667,234 (Duroc), 23,286,212 (Synthetic), and 25,709,552 (Large White) post-filtering variants, respectively. The average per variant call rate is above 98% for all breeds and the ratio transitions to transversions is between 2.33-2.35 (Table S4). Variant (SNPs, Indels) annotation was performed using the Variant Effect Predictor (VEP, release 97) [18].

4.6.7. pCADD scores

pCADD scores were retrieved from Gross et al. [11]. Visualization of pCADD scores was performed using JBrowse 1.16.6 [73]. Integration of sequence variants with pCADD score was performed using PyVCF [74]. pCADD scores, partitioned per chromosome, compressed via bgzip and tabix indexed for fast access, can be downloaded following this link (~5GB-1GB): http://www.bioinformatics.nl/pCADD/indexed_pPHRED-scores/, and scripts to use these scores to annotate SNPs can be found here: https://git.wur.nl/gross016/pcadd-scripts-data/.

4.6.8. Promoter and enhancer elements from ChipSeq data

We retrieved three H3K27Ac, and three H3K4me3 libraries (ArrayExpress accession number: E-MTAB-2633) from liver tissue from three male pig samples described by Villar et al. 2015. Data was aligned using BWA-mem [69] and visualized in JBrowse [73]. Coverage information on variant sites was obtained using PyVCF [74] and the PySAM 0.15.0 package.

4.6.9. Phenotypes and gene ontology

Phenotype information from genes orthologous to pig in human, mouse, and rat were retrieved from the Ensembl database (release 97) [75] using a custom bash script. Gene ontology and pathway information was obtained from the UniProt database [20].

4.6.10. RNA-sequencing and differential expression

We used 25 Duroc and 34 Landrace RNA-sequenced boars selected based on high and low sperm DNA fragmentation index, a measure of well packed double-stranded DNA vs single-stranded denatured DNA, which is an important indicator of boar fertility [22]. The boars were all born in the same period of time and a broad range of semen quality tests were conducted on ejaculates of these boars. Sequencing was done in two batches. Library preparation and sequencing strategy of the first batch can be found in van Son et al. 2017. The second batch was prepared using TruSeq mRNA stranded HT kit (Illumina) on a Sciclone NGSx liquid automation system (Perkin Elmer). A final library quality check was performed on a Fragment Analyser (Advanced Analytical Technologies, Inc) and by qPCR (Kapa Biosciences). Libraries were sequenced on an Illumina HiSeq 4000 according to manufacturer's instructions. Image analysis and base calling were performed using Illumina's RTA software v2.7.7. The resulting 100 basepair single-end reads were filtered for low base call quality using Illumina's default chastity criteria. We mapped the RNA-seq data to the Sscrofa11.1 reference genome using STAR [76] and called transcripts and normalized FPKM expression levels using Cufflinks and Cuffnorm [77]. We assigned the genotype class (homozygous reference, heterozygous, homozygous alternative) for each RNA-sequenced individual using the

660K genotype of the lead SNP in the GWAS result. We tested for differential expression between three genotype classes using the one-way ANOVA test. The Welch t-test was used to evaluate the differences between two genotype classes. A p value < 0.05 was considered significant.

Bibliography

[1] A. B. Gjuvsland, J. O. Vik, D. A. Beard, P. J. Hunter, and S. W. Omholt, "Bridging the genotype-phenotype gap: what does it take?," *J. Physiol.*, vol. 591, no. 8, pp. 2055–2066, 2013.

[2] D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping," *Nat. Rev. Genet.*, vol. 19, no. 8, pp. 491–504, 2018.

[3] M. D. Gallagher and A. S. Chen-Plotkin, "The post-GWAS era: from association to function," *Am. J. Hum. Genet.*, vol. 102, no. 5, pp. 717–730, 2018.

[4] C. P. Ponting and R. C. Hardison, "What fraction of the human genome is functional?," *Genome Res.*, vol. 21, no. 11, pp. 1769–1776, 2011.

[5] M. E. Cannon and K. L. Mohlke, "Deciphering the emerging complexities of molecular mechanisms at GWAS loci," *Am. J. Hum. Genet.*, vol. 103, no. 5, pp. 637–653, 2018.

[6] I. Dunham *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[7] S. J. G. Hall, "Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data," *Animal*, vol. 10, no. 11, pp. 1778–1785, 2016.

[8] R. Veroneze *et al.*, "Linkage disequilibrium and haplotype block structure in six commercial pig lines," *J. Anim. Sci.*, vol. 91, no. 8, pp. 3493–3501, 2013.

[9] J. N. Hellwege, J. M. Keaton, A. Giri, X. Gao, D. R. Velez Edwards, and T. L. Edwards, "Population stratification in genetic association studies," *Curr. Protoc. Hum. Genet.*, vol. 95, no. 1, pp. 1–22, 2017.

[10] M. Ron and J. I. Weller, "From QTL to QTN identification in livestock-winning by points rather than knock-out: a review," *Anim. Genet.*, vol. 38, no. 5, pp. 429–439, 2007.

[11] C. Groß *et al.*, "PCADD: SNV prioritisation in Sus scrofa," *Genet. Sel. Evol.*, vol. 52, no. 1, pp. 1–15, 2020.

[12] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: Predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, 2019.

[13] C. Groß, D. de Ridder, and M. Reinders, "Predicting variant deleteriousness in non-human species: Applying the CADD approach in mouse," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, 2018.

[14] E. F. Knol, B. Nielsen, and P. W. Knap, "Genomic selection in commercial pig breeding," *Anim. Front.*, vol. 6, no. 1, pp. 15–22, 2016.

[15] T. Meuwissen, B. Hayes, and M. Goddard, "Genomic selection: A paradigm shift in animal breeding," *Anim. Front.*, vol. 6, no. 1, pp. 6–14, 2016.

[16] D. Habier, R. L. Fernando, and D. J. Garrick, "Genomic BLUP decoded: a look into the black box of genomic prediction," *Genetics*, vol. 194, no. 3, pp. 597–607, 2013.

[17] M. E. Goddard, K. E. Kemper, I. M. MacLeod, A. J. Chamberlain, and B. J. Hayes, "Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture," *Proc. R. Soc. B Biol. Sci.*, vol. 283, no. 1835, p. 20160569, 2016.

[18] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor.," *Genome Biol.*, vol. 17, no. 1, p. 122, 2016.

[19] D. Villar *et al.*, "Enhancer evolution across 20 mammalian species," *Cell*, vol. 160, no. 3, pp. 554–566, 2015.

[20] U. Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506--D515, 2019.

[21] D. R. Zerbino *et al.*, "Ensembl 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, 2018.

[22] M. van Son *et al.*, "RNA sequencing reveals candidate genes and polymorphisms related to sperm DNA integrity in testis tissue from boars," *BMC Vet. Res.*, vol. 13, no. 1, p. 362, 2017.

[23] K. S. Kim, N. Larsen, T. Short, G. Plastow, and M. F. Rothschild, "A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits," *Mamm. genome*, vol. 11, no. 2, pp. 131–135, 2000.

[24] M. Van Son *et al.*, "A QTL for number of teats shows breed specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics," *Front. Genet.*, vol. 10, p. 272, 2019.

[25] D. Milan *et al.*, "A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle," *Science*, vol. 288, no. 5469, pp. 1248–1251, 2000.

[26] P. Uimari and A. Sironen, "A combination of two variants in PRKAG3is needed for a positive effect on meat quality in pigs," *BMC Genet.*, vol. 15, no. 1, p. 29, 2014.

[27] B. Essén-Gustavsson, A. Granlund, B. Benziane, M. Jensen-Waern, and A. V. Chibalin, "Muscle glycogen resynthesis, signalling and metabolic responses following acute exercise in exercise-trained pigs carrying the PRKAG3 mutation," *Exp. Physiol.*, vol. 96, no. 9, pp. 927–937, 2011.

[28] M. Mahlapuu *et al.*, "Expression profiling of the γ-subunit isoforms of AMP-activated protein kinase suggests a major role for γ3 in white skeletal muscle," *Am. J. Physiol. Metab.*, vol. 286, no. 2, pp. E194--E200, 2004.

[29] D. Ciobanu *et al.*, "Evidence for new alleles in the protein kinase adenosine monophosphateactivated γ3-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality," *Genetics*, vol. 159, no. 3, pp. 1151–1162, 2001.

[30] K. S. Kim *et al.*, "Association of melanocortin 4 receptor (MC4R) and high mobility group AT-hook 1 (HMGA1) polymorphisms with pig growth and fat deposition traits," *Anim. Genet.*, vol. 37, no. 4, pp. 419–421, 2006.

[31] J. Hong *et al.*, "Effects of genetic variants for the swine FABP3, HMGA1, MC4R, IGF2, and FABP4 genes on fatty acid composition," *Meat Sci.*, vol. 110, pp. 46–51, 2015.

[32] J. Chung *et al.*, "High mobility group A2 (HMGA2) deficiency in pigs leads to dwarfism, abnormal fetal resource

allocation, and cryptorchidism," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 21, pp. 5420–5425, 2018.
[33] A. Tanabe *et al.*, "Functional single-nucleotide polymorphisms in the secretogranin III (SCG3) gene that form secretory granules with appetite-related neuropeptides are associated with obesity," *J. Clin. Endocrinol. Metab.*, vol. 92, no. 3, pp. 1145–1154, 2007.
[34] J. A. Blake *et al.*, "Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D723–D729, 2017.
[35] S. Halpain and L. Dehmelt, "The MAP1 family of microtubule-associated proteins," *Genome Biol.*, vol. 7, no. 6, p. 224, 2006.
[36] C. M. Anderson, J. Hu, R. M. Barnes, A. B. Heidt, I. Cornelissen, and B. L. Black, "Myocyte enhancer factor 2C function in skeletal muscle is required for normal growth and glucose metabolism in mice," *Skelet. Muscle*, vol. 5, no. 1, pp. 1–10, 2015.
[37] S. Yoshizumi *et al.*, "Increased hepatic expression of ganglioside-specific sialidase, NEU3, improves insulin sensitivity and glucose tolerance in mice," *Metabolism*, vol. 56, no. 3, pp. 420–429, 2007.
[38] D. Sean Froese *et al.*, "Structural basis of glycogen branching enzyme deficiency and pharmacologic rescue by rational peptide design," *Hum. Mol. Genet.*, vol. 24, no. 20, pp. 5667–5676, 2015.
[39] W. Fan, F. Du, and X. Liu, "TRIM66 confers tumorigenicity of hepatocellular carcinoma cells by regulating GSK-3β-dependent Wnt/β-catenin signaling," *Eur. J. Pharmacol.*, vol. 850, no. 277, pp. 109–117, 2019.
[40] L. Nagy *et al.*, "Glycogen phosphorylase inhibition improves beta cell function," *Br. J. Pharmacol.*, vol. 175, no. 2, pp. 301–319, 2018.
[41] H. Baribault *et al.*, "The G-Protein-Coupled Receptor GPR103 Regulates Bone Formation," *Mol. Cell. Biol.*, vol. 26, no. 2, pp. 709–717, 2006.
[42] J. Sjölund, F. G. Pelorosso, D. A. Quigley, R. DelRosario, and A. Balmain, "Identification of Hipk2 as an essential regulator of white fat development," *Proc. Natl. Acad. Sci.*, vol. 111, no. 20, pp. 7373–7378, 2014.
[43] H. Reyer, S. Ponsuksili, K. Wimmers, and E. Murani, "Transcript variants of the porcine glucocorticoid receptor gene (NR3C1)," *Gen. Comp. Endocrinol.*, vol. 189, pp. 127–133, 2013.
[44] S. Casiró *et al.*, "Genome-Wide association study in an F2 duroc x pietrain resource population for economically important meat quality and carcass traits," *J. Anim. Sci.*, vol. 95, no. 2, pp. 545–558, 2017.
[45] J. Yun *et al.*, "RNA-Seq analysis reveals a positive role of HTR2A in adipogenesis in Yan Yellow Cattle," *Int. J. Mol. Sci.*, vol. 19, no. 6, p. 1760, 2018.
[46] Y. Liu, W. Qi, A. Richardson, H. Van Remmen, Y. Ikeno, and A. B. Salmon, "Oxidative damage associated with obesity is prevented by overexpression of CuZn-or Mn-superoxide dismutase," *Biochem. Biophys. Res. Commun.*, vol. 438, no. 1, pp. 78–83, 2013.
[47] A. Castrillo, D. J. Pennington, F. Otto, P. J. Parker, M. J. Owen, and L. Boscá, "Protein kinase Cε is required for macrophage activation and defense against bacterial infection," *J. Exp. Med.*, vol. 194, no. 9, pp. 1231–1242, 2001.
[48] M. J. Jang, U.-H. Park, J. W. Kim, H. Choi, S.-J. Um, and E.-J. Kim, "CACUL1 reciprocally regulates SIRT1 and LSD1 to repress PPARγ and inhibit adipogenesis," *Cell Death Dis.*, vol. 8, no. 12, pp. 1–14, 2017.
[49] M. Niwa *et al.*, "IRAP deficiency attenuates diet-induced obesity in mice through increased energy expenditure," *Biochem. Biophys. Res. Commun.*, vol. 457, no. 1, pp. 12–18, 2015.
[50] L. J. Den Hartigh *et al.*, "Deletion of serum amyloid A3 improves high fat high sucrose diet-induced adipose tissue inflammation and hyperlipidemia in female mice," *PLoS One*, vol. 9, no. 9, pp. 1–13, 2014.
[51] K. Mu *et al.*, "Hepatic nitric oxide synthase 1 adaptor protein regulates glucose homeostasis and hepatic insulin sensitivity in obese mice depending on its PDZ binding domain," *EBioMedicine*, vol. 47, pp. 352–364, 2019.
[52] K. Piórkowska, K. Ropka-Molik, T. Szmatoła, K. Zygmunt, and M. Tyra, "Association of a new mobile element in predicted promoter region of ATP-binding cassette transporter 12 gene (ABCA12) with pig production traits," *Livest. Sci.*, vol. 168, pp. 38–44, 2014.
[53] X. Li *et al.*, "Investigation of porcine FABP3 and LEPR gene polymorphisms and mRNA expression for variation in intramuscular fat content," *Mol. Biol. Rep.*, vol. 37, no. 8, pp. 3931–3939, 2010.
[54] A. Adhikari, P. Mainali, and J. K. Davie, "JARID2 and the PRC2 complex regulate the cell cycle in skeletal muscle," *J. Biol. Chem.*, vol. 294, no. 51, pp. 1–14, 2019.
[55] B. Zhang, P. Shang, Y. Qiangba, A. Xu, Z. Wang, and H. Zhang, "The association of NR1H3 gene with lipid deposition in the pig," *Lipids Health Dis.*, vol. 15, no. 1, p. 99, 2016.
[56] C. J. Sinal, M. Tohkin, M. Miyata, J. M. Ward, G. Lambert, and F. J. Gonzalez, "Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis," *Cell*, vol. 102, no. 6, pp. 731–744, 2000.
[57] R. D. Palmiter, "Reduced levels of neurotransmitter-degrading enzyme PRCP promote obesity," *J. Clin. Invest.*, vol. 119, no. 8, pp. 2130–2133, 2009.
[58] J. Trevaskis *et al.*, "Src homology 3-domain growth factor receptor-bound 2-like (endophilin) interacting protein 1, a novel neuronal protein that regulates energy balance," *Endocrinology*, vol. 146, no. 9, pp. 3757–3764, 2005.
[59] K. Rosenvold *et al.*, "Muscle glycogen stores and meat quality as affected by strategic finishing feeding of slaughter pigs," *J. Anim. Sci.*, vol. 79, no. 2, pp. 382–391, 2001.
[60] E. Giuffra and C. K. Tuggle, "Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap," *Annu. Rev. Anim. Biosci.*, vol. 7, no. 1, pp. 65–88, 2019.
[61] M. Georges, C. Charlier, and B. Hayes, "Harnessing genomic information for livestock improvement," *Nat. Rev. Genet.*, vol. 20, no. 3, pp. 135–156, 2019.

[62] Y. Bao, P. Kishnani, J. Y. Wu, and Y. T. Chen, "Hepatic and neuromuscular forms of glycogen storage disease type IV caused by mutations in the same glycogen-branching enzyme gene," *J. Clin. Invest.*, vol. 97, no. 4, pp. 941–948, 1996.
[63] C. Perleberg, A. Kind, and A. Schnieke, "Genetically engineered pigs as models for human disease," *Dis. Model. Mech.*, vol. 11, no. 1, 2018.
[64] A. Warr *et al.*, "An improved pig reference genome sequence to enable pig genetics and genomics research," *bioRxiv*, p. 668921, 2019.
[65] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.
[66] Q. Mary, M. End, and C. Biology, "ASReml User Guide," 2009.
[67] D. J. Garrick, J. F. Taylor, and R. L. Fernando, "Deregressing estimated breeding values and weighting information for genomic regression analyses," *Genet. Sel. Evol.*, vol. 41, no. 1, pp. 1–8, 2009.
[68] "FastQC: a quality control tool for high throughput sequence data." Babraham Institute, Cambridge UK, 2011.
[69] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.
[70] G. G. Faust and I. M. Hall, "SAMBLASTER: Fast duplicate marking and structural variant read extraction," *Bioinformatics*, vol. 30, no. 17, pp. 2503–2505, 2014.
[71] H. Li *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.
[72] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," pp. 1–9, 2012.
[73] R. Buels *et al.*, "JBrowse : a dynamic web platform for genome visualization and analysis," *Genome Biol.*, vol. 17, no. 66, pp. 1–12, 2016.
[74] J. Casbon, "PyVCF - A variant call format parser for Python." 2012.
[75] S. E. Hunt *et al.*, "Ensembl variation resources," *Database*, vol. 2018, pp. 1–12, 2018.
[76] A. Dobin *et al.*, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.
[77] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nat. Biotechnol.*, vol. 31, no. 1, p. 46, 2013.

4.7. Appendix - Supplementary data

4.7.1. Supplementary note

4.7.1.1. A missense mutation in the MC4R gene affects production traits in the Synthetic and Duroc breed.

A QTL on SSC1 affects lifetime growth rate, backfat and feed intake (Figure S1) in the Synthetic and Duroc breed. The lead SNP in Duroc is located at 1:g.159884741A>C, and the C allele is associated with less backfat but also slower growth (AF=44%). The lead SNP in the Synthetic breed is located at position 1:g.159660303C>T, and the T allele is associated with less backfat and slower growth (AF=11%). The analysis reveals 526 and 315 variants in high LD with the lead SNP in the Synthetic and Duroc breed, respectively. A missense variant (1:g.160773437G>A) in the *MC4R* gene shows the highest pCADD score in both variant sets (score=27.48). The variant substitutes aspartic acid for an asparagine (ENSSSCG00000004904:p.Asp298Asn), segregating at a frequency of approximately 88%, and 55% in Synthetic and Duroc, respectively. This gene is well-known to affect obesity and fatness traits and the missense variant has been reported in various pig breeds [23], [30], [31]. The Asp298 allele is associated with less backfat, slower growth, and lower feed intake, while the Asn298 allele is associated with more fat, higher-feed consumption, and faster growth.

4.7.1.2. A promoter mutation in the VRTN gene affects number of vertebrae and number of teats.

A QTL on SSC7 affects the number of teats and the number of vertebrae in Landrace, Duroc and Large White (Figure S3). The lead SNPs for all three breeds are found within a 100kb region (Duroc: 7:g.97614635A>G, Landrace: 7:g.97652632T>C, Large White: 7:g.97568284A>G). The analysis reveals 526, 6,553 and 315 variants in high LD with the lead SNP in the Duroc, Landrace and Large White breed, respectively. Two variants affecting *VRTN* expression are proven to increase the number of vertebrae, and thereby also the number of teats [24]. One variant affects the *VRTN* promoter (7:g. 97614602A>G) and the other is a PRE1 (7:97615896_ins291) insertion element in the first intron of the *VRTN* gene. The promoter variant is one of the top pCADD scored variants in Duroc (2nd) and Large White (4th), while the variant is not among the top variants in Landrace due to the large LD block covering 6,553 variants. The PRE1 element is not included because we cannot infer pCADD scores for such type of variants. This example shows that the tool can be very powerful to prioritize variants, but with a trade-off for the level of LD, increasing the noise if many thousands of variants are in linkage.

4.7.2. Supplementary figures

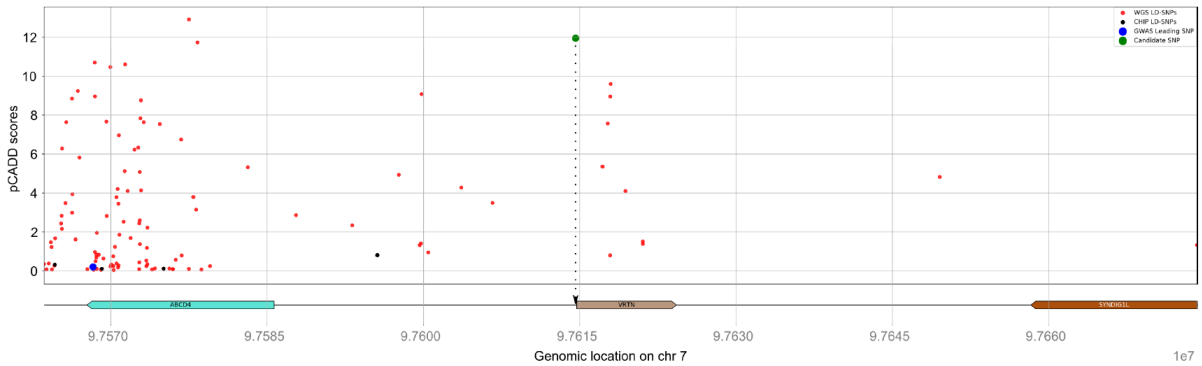


Figure S1: Plot showing all sequence variants in high LD (red) with the lead SNP of the SSC7 QTL for number of teats (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98.

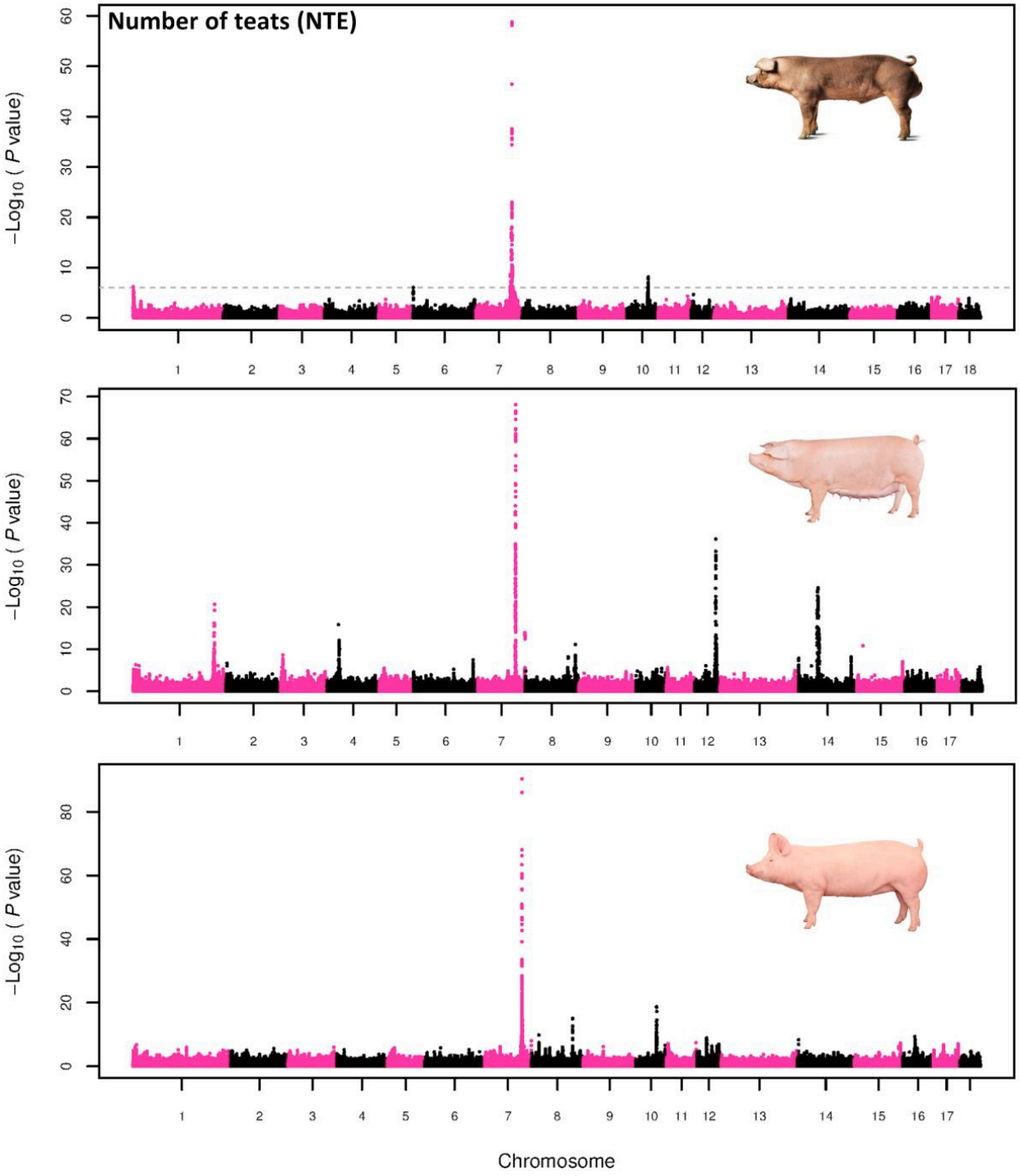


Figure S2: Manhattan plot for number of teats in Duroc, Landrace, and Large White. A strong QTL on chromosome 7 is observed across the breeds.

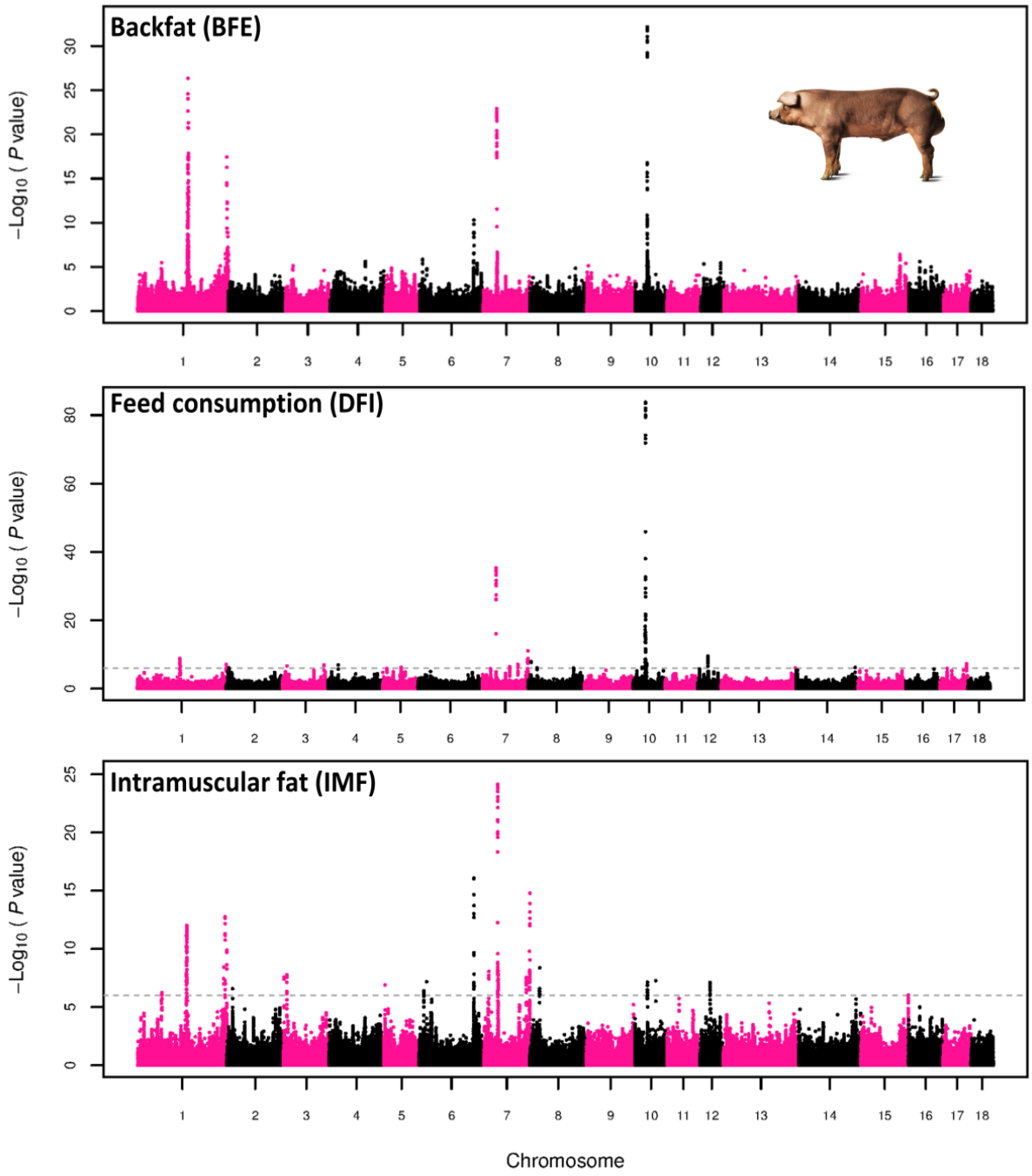


Figure S3: Manhattan plot for backfat, daily gain, and intramuscular fat in the Duroc breed. A single strong QTL on SSC7 affects all three traits.

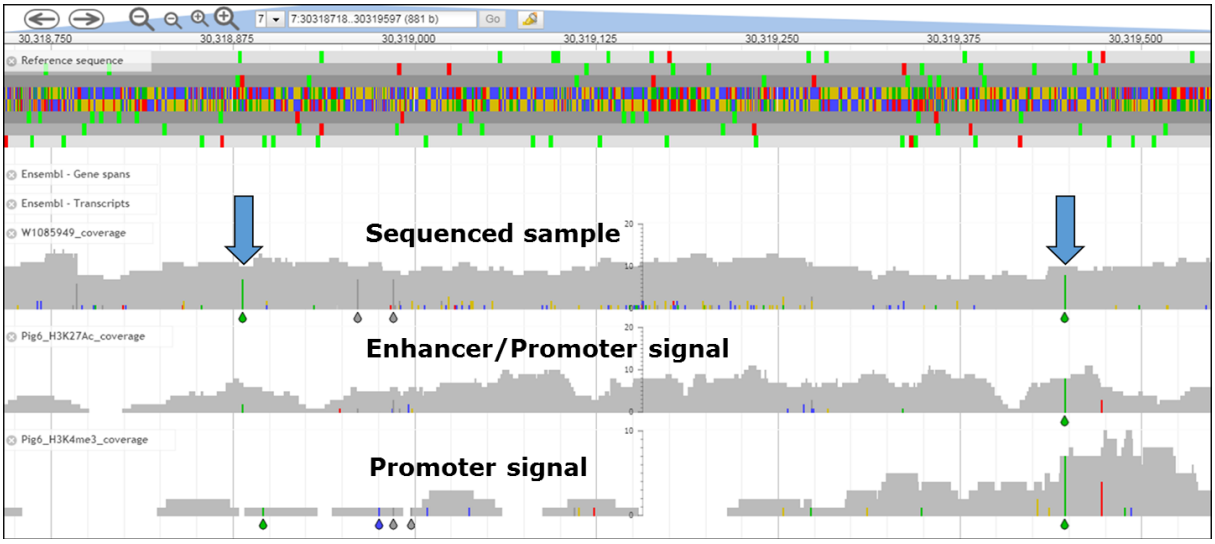


Figure S4: JBrowse screen capture showing the coverage track of one heterozygous sample for the two HMGA1 upstream gene variants (indicated with the blue arrows). The variants overlap with the promotor region of the gene, supported by signals on the H3K4me3 and H3K27Ac in liver tissue [19].

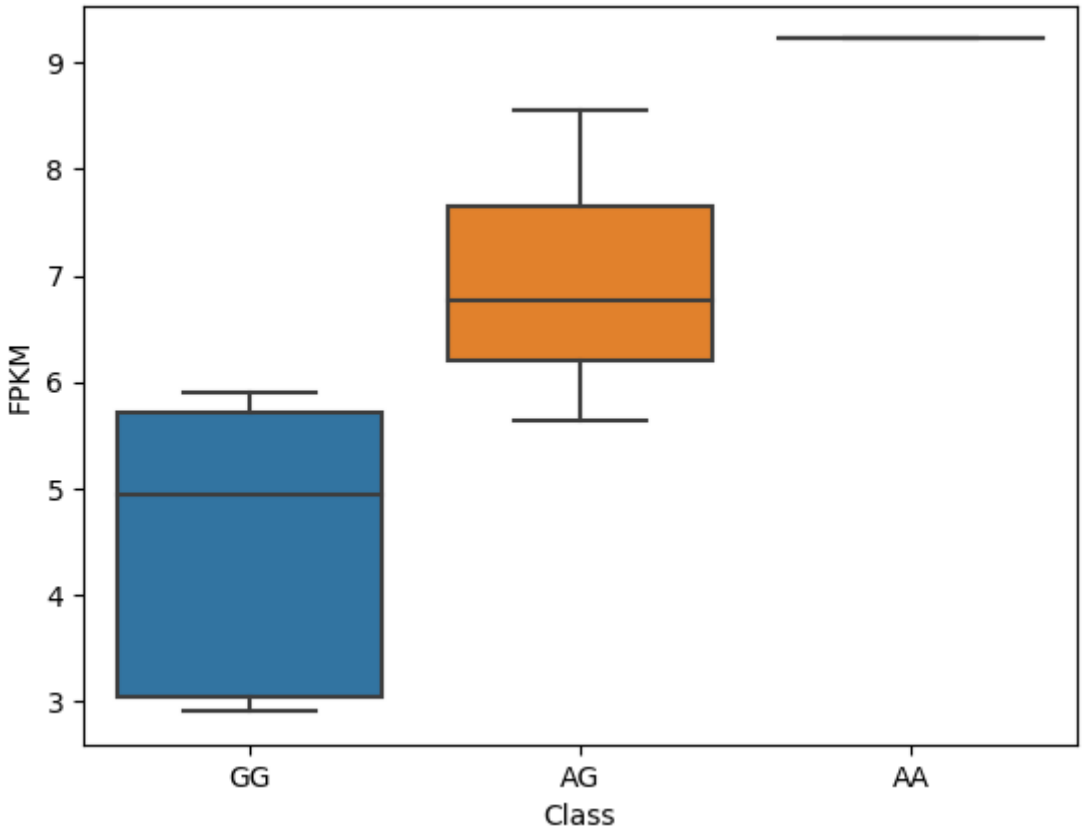


Figure S5: FPKM expression values for different classes of genotypes for the HMGA1 7:g.30319447G>A variant. HMGA1 expression increases additively for the A allele.

4.7.3. Supplementary table

Table S1: Statistics and percentiles of pCADD score per variant effect predictor (VEP) class.

Abbr.	SL	SG	CS	NC	U3	S	IG	NS	SN	U5	UP	I	DN
Consequence	Stop Lost	Stop Gained	Canonical Splice	Noncoding Change	3' UTR	Splice Site	Intergenic	Non-Synonymous	Synonymous	5' UTR	Upstream	Intronic	Downstream
# Variants	102	776	724	9386	154014	15769	10817716	73855	197655	29617	814663	7889447	722111
Min	0	0	0	0	0	0	0	0	0	0	0	0	0
1-Percentile	0.16	0	0	0.03	0.03	0	0.03	0	0	0.02	0.06	0.04	0.04
5-Percentile	13.8	0	0.06	0.17	0.14	0.03	0.2	0	0.01	0.14	0.28	0.18	0.24
10-Percentile	20.6	0.02	1.12	0.42	0.3	0.09	0.37	0	0.04	0.27	0.53	0.35	0.46
25-Percentile	26.2	8.31	4.97	1.22	0.86	0.44	1.11	0.07	0.15	0.84	1.33	1.01	1.22
50-Percentile	35.5	19.2	10.19	2.65	2.04	1.57	2.7	10.55	0.53	3.46	2.66	2.53	2.67
75-Percentile	40.1	23.7	14.96	4.81	4.33	5.12	5.52	17.9	2.1	8.91	4.54	5.32	5.02
90-Percentile	44.5	30.5	23.44	8.02	7.91	11.31	8.7	22.13	6.39	12.61	6.87	8.59	7.85
95-Percentile	46.7	36.1	27.95	10.79	11.5	13.37	10.25	23.9	9.88	14.7	8.59	10.2	9.65
99-Percentile	52.1	43.2	33.97	15.45	15.78	16.74	14.01	27.96	13.51	19.01	13.88	14.01	14.04
Max	54.6	51.2	39.45	24.97	28.68	24.19	29.68	66.81	29.96	33.36	37.48	39.65	39.17
Std	10.7	11.2	8.24	3.38	3.53	4.43	3.34	9.12	3.14	5.03	2.85	3.34	3.13

Table S2: RNA-sequenced samples alignment and breed information

RNAseq_ID	Birthdate	Breed	# Aligned reads	# Aligned to genes
200982	19.06.2009	Landrace	81,072,871	50,707,721
201267	30.06.2009	Landrace	77,004,472	45,680,563
201839	20.07.2009	Landrace	80,925,045	49,181,261
202393	04.08.2009	Landrace	77,882,188	47,046,576
202517	25.08.2009	Duroc	99,035,646	58,171,426
202553	02.09.2009	Duroc	67,649,946	40,056,978
202889	08.09.2009	Landrace	83,764,698	53,617,685
202956	22.08.2009	Landrace	79,990,131	50,997,185
203757	07.10.2009	Landrace	88,166,223	53,139,698
204120	30.10.2009	Landrace	85,272,052	53,810,459
211478	13.08.2010	Duroc	88,652,551	52,520,221
212051	29.08.2010	Landrace	93,711,956	54,912,183
214807	16.12.2010	Landrace	70,420,072	42,079,270
216352	23.02.2011	Duroc	77,037,167	44,240,695
217000	25.02.2011	Duroc	85,948,543	50,676,113
219433	03.06.2011	Duroc	78,895,825	46,292,107
220011	26.06.2011	Duroc	68,126,200	40,688,175
220166	25.06.2011	Duroc	90,856,057	53,930,696
220199	09.07.2011	Landrace	87,849,053	53,145,870
221522	13.08.2011	Duroc	76,362,846	46,013,378
222375	07.09.2011	Duroc	70,498,201	42,717,332
222726	27.09.2011	Landrace	97,524,851	58,333,275
223984	17.11.2011	Duroc	76,542,504	44,902,014
223989	16.11.2011	Duroc	91,391,099	54,432,619
224059	29.11.2011	Duroc	66,451,777	38,339,124
224396	20.11.2011	Landrace	103,566,935	63,129,636
225577	27.12.2011	Duroc	84,992,811	49,850,466
226260	18.01.2012	Duroc	74,843,271	43,972,824
229039	04.05.2012	Duroc	88,647,993	53,712,798
230319	26.06.2012	Landrace	86,080,717	50,725,416
601979	09.07.2013	Landrace	116,786,753	67,559,389
606695	21.07.2013	Landrace	143,196,072	87,965,328
618265	02.11.2011	Landrace	97,827,095	57,921,236
618292	10.12.2012	Landrace	131,790,646	82,814,520
618315	01.06.2013	Duroc	78,829,116	50,107,043
632533	14.11.2013	Landrace	144,901,184	82,206,500
632566	01.10.2013	Landrace	85,825,439	53,844,726
652024	02.11.2013	Landrace	140,421,410	87,491,329
662985	18.12.2013	Landrace	87,012,842	51,000,123
667317	19.12.2013	Duroc	89,667,304	54,426,898
671867	01.08.2013	Landrace	82,551,158	49,932,527
671868	18.02.2014	Landrace	91,297,409	52,777,850
671879	03.08.2013	Landrace	124,472,518	74,334,920
678115	03.09.2013	Landrace	96,432,147	58,449,186

679241	18.10.2013	Landrace	81,233,028	48,725,698
679906	05.07.2013	Duroc	77,563,765	47,396,856
679941	02.06.2013	Duroc	202,290,492	113,380,344
679943	18.01.2014	Landrace	74,056,318	43,376,629
687220	17.09.2013	Landrace	81,889,922	45,368,799
716309	26.07.2013	Duroc	67,791,177	37,218,720
721795	05.05.2014	Landrace	87,454,783	52,381,283
740620	12.05.2014	Landrace	80,399,150	48,285,099
761156	05.05.2014	Landrace	86,663,844	52,182,153
765489	18.11.2013	Duroc	75,123,881	42,499,024
775352	11.04.2014	Duroc	124,247,297	76,413,078
787897	13.12.2013	Duroc	80,394,861	49,831,000
793958	30.01.2014	Duroc	71,248,334	39,648,673
793985	20.12.2013	Landrace	124,510,805	68,625,964
822998	17.12.2013	Landrace	111,917,137	69,986,352

Table S3: Gene ontology enrichment analysis for genes underlying important selection traits in pigs.

Enrichment FDR	Genes in list	Total genes	Functional Category
1.80E-05	15	2004	Homeostatic process
1.80E-05	9	516	Multicellular organismal homeostasis
5.00E-05	5	91	Energy reserve metabolic process
5.00E-05	9	627	Behavior
3.20E-04	2	2	Oncogene-induced cell senescence
3.20E-04	5	147	Positive regulation of lipid metabolic process
3.20E-04	5	148	Positive regulation of small molecule metabolic process
3.20E-04	7	438	Muscle system process
3.70E-04	4	72	Regulation of behavior
3.70E-04	7	463	Regulation of small molecule metabolic process
3.90E-04	4	78	Glycogen metabolic process
3.90E-04	4	79	Cellular glucan metabolic process
3.90E-04	4	79	Glucan metabolic process
4.10E-04	5	183	Organ growth
4.10E-04	2	3	Senescence-associated heterochromatin focus assembly
4.10E-04	4	84	Positive regulation of lipid biosynthetic process
4.10E-04	3	26	Regulation of feeding behavior
4.20E-04	5	190	Regulation of lipid biosynthetic process
4.40E-04	8	742	Cellular response to hormone stimulus
6.00E-04	9	1040	Response to hormone
6.90E-04	11	1681	Negative regulation of biosynthetic process
6.90E-04	4	104	Cellular polysaccharide metabolic process
7.20E-04	11	1704	Response to endogenous stimulus
8.00E-04	3	36	Energy homeostasis
8.30E-04	4	116	Polysaccharide metabolic process
8.30E-04	6	401	Regulation of lipid metabolic process
8.30E-04	3	38	Toll-like receptor 4 signaling pathway
8.30E-04	10	1432	Cellular response to endogenous stimulus
9.20E-04	15	3382	Regulation of multicellular organismal process
1.10E-03	7	644	Carbohydrate metabolic process

Table S4: Population whole genome sequencing statistics

Breed	# Samples	# SNPs	# Indels	Avg. per variant call rate	Ts/Tv
Landrace	167	17,899,503	3,759,040	98.85	2.33
Duroc	119	19,696,060	3,946,140	98.92	2.34
Synthetic	71	19,435,050	3,832,091	98.01	2.34
Large White	89	25,709,552	4,085,600	98.46	2.35

5. Prioritizing sequence variants in conserved non-coding elements in the chicken genome using chCADD

Christian Groß*

Chiara Bortoluzzi*

Dick de Ridder

Hendrik-Jan Megens

Martien A.M. Groenen

Marcel Reinders

Mirte Bosse

*shared first authorship

This chapter is published in bioRxiv and accepted (with the above title) by PLOS Genetics.

C. Groß, C. Bortoluzzi, D. de Ridder, et al., "Evolutionarily conserved non-protein-coding regions in the chicken genome harbor functionally important variation", *bioRxiv* 2020.03.27.012005; doi: https://doi.org/10.1101/2020.03.27.012005

5.1. Abstract

The availability of genomes for many species has advanced our understanding of the non-protein-coding fraction of the genome. Comparative genomics has proven to be an invaluable approach for the systematic, genome-wide identification of conserved non-protein-coding elements (CNEs). However, for many non-mammalian model species, including chicken, our capability to interpret the functional importance of variants overlapping CNEs has been limited by current genomic annotations, which rely on a single information type (e.g. conservation). We here studied CNEs in chicken using a combination of population genomics and comparative genomics. To investigate the functional importance of variants found in CNEs we develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD), a variant effect prediction tool first introduced for humans and later on for mouse and pig. We show that 73 Mb of the chicken genome has been conserved across more than 280 million years of vertebrate evolution. The vast majority of the conserved elements are in non-protein-coding regions, which display SNP densities and allele frequency distributions characteristic of genomic regions constrained by purifying selection. By annotating SNPs with the chCADD score we are able to pinpoint specific subregions of the CNEs to be of higher functional importance, as supported by SNPs found in these subregions are associated with known disease genes in humans, mice, and rats. Taken together, our findings indicate that CNEs harbor variants of functional significance that should be object of further investigation along with protein-coding mutations. We therefore anticipate chCADD to be of great use to the scientific community and breeding companies in future functional studies in chicken.

5.2. Introduction

The rapidly increasing availability of genomes has considerably advanced our understanding of the non-protein-coding fraction of the genome. With the sequencing of the human genome [1] and the first ENCODE project [2], [3] it was soon realized that protein-coding genes constitute a small fraction of a species functional genome and that the remaining non-protein-coding DNA is not simply ´junk´ DNA as initially thought. Nevertheless, the functional importance of these non-protein-coding regions remained for long time unknown, as determining (molecular) function was far more difficult than for protein-coding genes [4]. A better understanding of the functional importance of these non-protein-coding regions comes from comparative genomics, which has allowed the systematic, genome-wide identification of conserved non-protein-coding elements (CNEs) [5], [6].

Comparative genomics relies on the genome comparison of a group of species related by a narrow or wide time-scale (i.e. phylogenetic scope). Regions in the genome that share some minimum sequence similarity across two or more species are an indication of a selection constraint. Moreover, conservation often implies a biological function [7]. Based on this principle, CNEs can be identified in any species included in the alignment, as reported in recent studies in the collared flycatcher [8], fruit flies [9], and plants [6]. However, the phylogenetic scope [10] and species included in the alignment [11] can have important implications for the identification of CNEs. For instance, by including the spotted gar genome in their alignment, [11] recently identified numerous CNEs previously undetectable in direct human-teleost comparisons, supporting the importance of a bridging species in the alignment.

CNEs have been the subject of intense recent interest. The identification of CNEs has had important implications in enhancing genome annotation [12], investigating signatures of adaptive evolution [13]–[15], and identifying putative trait loci [16]. CNEs and sequence conservation have also proven crucial in studying the genetic basis of phenotypic diversity. In fact, non-protein-coding SNPs have been linked to traits and diseases in genome-wide association studies [17], [18].

Although the methodological advantages of a comparative genomic approach are well recognized, the functional interpretation of CNEs is incomplete if based on conservation alone, as conservation provides information on restrictions, but not on functionality. A possible solution is combining conservation with other complementary types of data that characterize the biological role of genetic sequences at a genome-wide scale [7]. Such data include, for instance, RNA sequencing (RNA-seq) for the identification of transcriptionally active regions [19] and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for regulatory-factor-binding regions (RFBRs) [20]. In human genetics, integrative annotations such as Combined Annotation-Dependent Depletion (CADD) [21] have been developed. The main advantage of such frameworks is the combination, into a unique score, of diverse genomic features derived from, among others, gene model annotations, evolutionary constraints, epigenetic measurements, and functional predictions [21], [22].

Compared to humans, for many non-mammalian model species, including chicken (Gallus gallus), the situation is quite different. First, comparative genomic studies that made use of the very first genome assemblies [23]–[25] may have provided an incomplete and biased picture of avian CNEs and avian genome evolution, as recently pointed out by [26]. Second, the lack of species-specific methods that can identify and score functional non-protein-coding mutations throughout the genome has restricted most of the research interest to protein-coding genes. In fact, in the context of protein-coding genes generic predictors such as SIFT [27], PolyPhen2 [28], and Provean [29] can be used.

We here addressed these limitations using a combination of comparative genomic and population genomic approaches to accurately predict CNEs in the chicken genome. Furthermore, we used machine learning to develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD), in the tradition of previous CADD models for non-human species, including mouse (mCADD) [30] and pig (pCADD) [31]. As we show, chCADD has the potential of providing new insights into the functional role of non-protein-coding regions of the chicken genome at a single base pair resolution.

Even though deciphering the function of the non-protein-coding portion of a species genome has been a challenging task, we expect our study to provide a new framework for decoding the still largely unknown function of CNEs and their relative variants in chicken, an ideal non-mammalian model and anchor species in evolutionary studies.

5.3. Results

5.3.1. Conserved non-protein-coding elements cover a large fraction of the chicken genome

To define CNEs, we first identified conserved elements (CEs) using the UCSC PhastCons most conserved track approach [32]. PhastCons predicted in the 23 sauropsids multiple sequence alignment (MSA) 1.14 million CEs encompassing ~8% of the chicken genome for a total of 73 Mb. In line with the density of genes and regulatory features characteristic of the chicken genome [33], we found that most of the predicted CEs are on micro-chromosomes (GGA11-GGA33), followed by intermediate (GGA6-GGA10) and macro-chromosomes (GGA1-GGA5) (Figure S1). Even though the length of predicted CEs ranged from 4 bp to a maximum of ~ 2,000 bp, the vast majority was short (< 100 bp) (Figure S2). Therefore, we do not expect any length bias in our final set of CEs.

We annotated CEs by genomic features, considering only genes for which the transcript had a proper annotated start and stop codon, as defined by the Ensembl´s annotation files (n = 14,828 genes). Overall, we found that 23% of the predicted CEs were associated with exonic sequences

Table 1: Statistics of predicted conserved elements (CEs) based by gene annotations. The fraction of CEs per sites class is presented, for protein-coding gene annotations, in percentages of the exonic CEs (17,148,879 bp). For non-protein-coding gene annotations, the fraction is relative to the non-exonic CEs (51,224,645 bp). Abbreviations: CC, conserved coding; CNE, conserved non-protein-coding elements.

Genomic feature	No. overlapping CEs	Total overlap (bp)	Genome coverage (%)	Fraction of site class conserved (%)
CDS	213,787	14,683,183	1.38	85.62
5′ UTRs	5,457	207,320	0.02	1.21
3′ UTRs	23,721	1,460,144	0.15	8.51
Promoters	16,022	761,504	0.08	4.44
RNA genes	701	36,728	0.00	0.21
LncRNAs	121,840	7,696,557	0.80	15.03
Introns	328,579	18,520,675	1.93	36.16
Intergenic	400,501	25,007,413	2.60	48.82
Total CC	**259,688**	**17,148,879**	**1.78**	**100.00**
Total CNE	**850,920**	**51,224,645**	**5.33**	**100.00**

(i.e. CDS, 5′ UTR, 3′ UTR, promoter, and RNA genes) spanning 17.14 Mb of the chicken genome (Table 1). The majority of the exon-associated CEs overlapped known coding regions (85% of total exon-associated CEs), followed by 3′ UTRs (8% of total), and promoter regions (4% of total). Although we observed conservation in exon sequences, most CEs overlapped non-protein-coding sequences, including lncRNA (15% of total non-exon associated CEs), intronic (36% of total), and intergenic regions (49% of total). We further examined the biological processes and molecular functions of known genes overlapped by CEs in coding regions, 5′ UTRs, 3′ UTRs, and introns. These genes are associated with basic functions, including cell differentiation and development, anatomical structure development, morphogenesis, and growth (Table 2). Most of these GO categories have also been previously associated with mammalian and vertebrate ultraconserved elements (UCEs) [33], [34].

In total we identified 259,688 CEs in protein-coding regions, leaving 850,920 CNEs spanning over 51 Mb of the chicken genome (Table 1), with a genome-wide distribution of 92.10 CNEs/100-kb. We further observed noticeable differences in the length distribution of CEs associated with different types of annotations. Among the conserved exon-associated CEs, those found in CDSs are, on average, the longest (~68 bp), followed by 3′ UTRs (61 bp), RNA genes (52 bp), promoters (47 bp), and 5′ UTRs (38 bp) (Figure S3). On the contrary, CEs found in non-protein-coding regions show a homogenous length distribution, ranging from 56 bp in introns to 63 bp in lncRNAs (Figure S4).

Table 2: GO term enrichment analysis of exonic-associated CE and intronic CEs

Term ID	Term description	Target size	3 UTR				Intron			
			Term size	Query size	Overlap size	p-value	Term size	Query size	Overlap size	p-value
GO:0048856	Anatomical structure development	12,514	3,293	4,736	1,475	1.24×10^{-17}	3,293	6,971	2,128	1.09×10^{-29}
GO:0010646	Regulation of cell communication	12,514	2,038	4,736	917	3.67×10^{-09}	2,038	6,971	1,329	1.33×10^{-17}
GO:0010604	Positive regulation of macromolecule metabolic process	12,514	2,118	4,736	952	1.49×10^{-09}	2,118	6,971	1,331	2.21×10^{-09}
GO:0023051	Regulating of signaling	12,514	2,056	4,736	926	2×10^{-09}	2,056	6,971	1,339	1.88×10^{-17}
GO:0048583	Regulation of response to stimulus	12,514	2,332	4,736	1,032	1.44×10^{-08}	2,332	6,971	1,477	9.79×10^{-13}
GO:0048468	Cell development	12,514	1,364	4,736	625	1.27×10^{-06}	1,364	6,971	927	1.12×10^{-18}
GO:0031325	Positive regulation of cellular metabolic process	12,514	2,091	4,736	936	9.01×10^{-09}	2,091	6,971	1,304	1.09×10^{-07}

Term ID	Term description	Target size	CDS				5 UTR			
			Term size	Query size	Overlap size	p-value	Term size	Query size	Overlap size	p-value
GO:0048856	Anatomical structure development	12,514	3,293	9,703	2,713	2.06×10^{-11}	3,293	1,896	654	5.13×10^{-14}
GO:0010646	Regulation of cell communication	12,514	2,038	9,703	1,686	2.64×10^{-06}	2,038	1,896	381	9.33×10^{-03}
GO:0010604	Positive regulation of macromolecule metabolic process	12,514	2,118	9,703	1,749	3.53×10^{-06}	2,118	1,896	403	5.06×10^{-04}
GO:0023051	Regulating of signaling	12,514	2,056	9,703	1,699	4.46×10^{-06}	2,056	1,896	384	9.24×10^{-03}
GO:0048583	Regulation of response to stimulus	12,514	2,332	9,703	1918	5.55×10^{-06}	2,332	1,896	424	4.39×10^{-02}
GO:0048468	Cell development	12,514	1,364	9,703	1,142	1.78×10^{-05}	1,364	1,896	282	3.38×10^{-05}
GO:0031325	Positive regulation of cellular metabolic process	12,514	2,091	9,703	1,723	1.91×10^{-05}	2,091	1,896	388	1.60×10^{-02}

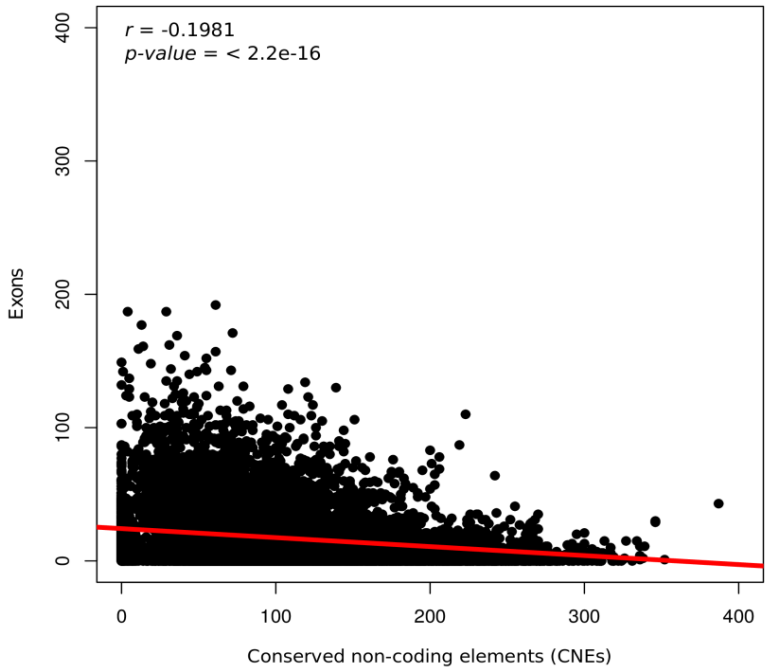


Figure 1: Correlation between exons and conserved non-protein-coding elements (CNEs) along the chicken genome. CNEs and exons count per 100 kb windows are shown with the Pearson correlation coefficient r and corresponding p-value in the top left corner.

5.3.2. CNEs populate regions not occupied by genes

We further investigated the genomic location of CNEs as this might provide important clues to their functional role. We found that the distribution of CNEs in windows of 100 kb is significantly negatively correlated (r = -0.20; p-value: <2.2x10-16) with the distribution of exons (Figure 1). We subsequently analyzed chicken polymorphism data to address the mutational or evolutionary forces shaping CNEs, following previous studies in humans [35] and Drosophila [9], [36]. We used polymorphism densities to investigate whether these forces could still be acting on the chicken genome or they could have acted in other species and may no longer be relevant for chicken. SNP density, which reflects events within the chicken lineage, was calculated in the genomes of 169 chickens from different traditional breeds of divergent demographic and selection history. Specifically, we compared the SNP density found in CNEs with that in non-protein-coding elements that were identified not to be conserved (non-CNEs; i.e. not conserved intronic, lncRNA and intergenic regions), following [9], [35], [36]. Overall, we found that CNEs are less enriched in SNPs (SNP density = 0.0092) than non-CNEs (SNP density = 0.02).

5.3.3. CNEs are selectively constrained in chicken

To test whether low local mutation rates in CNEs or purifying selection is responsible for the observed low SNP density, we looked at the derived allele frequency (DAF) distribution in CNEs and non-CNEs. This is because mutation rate differences are not expected to affect the allele frequency spectra. On the contrary, selective constraint is responsible for the shift in allele frequency distribution of constrained alleles towards lower values. Allele frequencies for derived (new) alleles were compiled using the sequence of the inferred ancestor between chicken and turkey. The ancestral allele was determined for a total of ~9 million SNPs that passed several filtering criteria

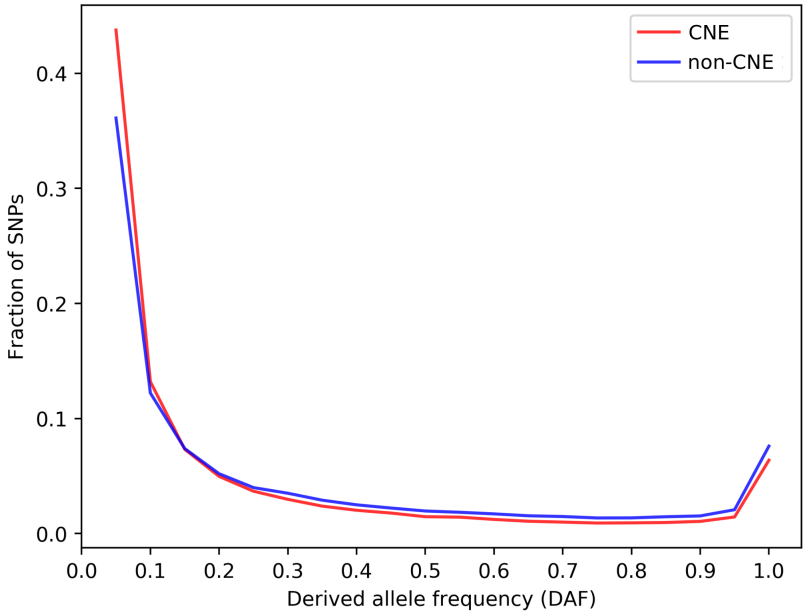


Figure 2: Derived allele frequency (DAF) distribution of SNPs in CNEs and non-CNEs.

(see Methods). We observed an excess of rare (≤ 10%) derived alleles of SNPs within CNEs in all chicken populations (Figure 2). Overall, 57% of SNPs within CNEs had a DAF ≤ 10%, compared to only ~48% in non-CNEs (the same pattern was observed for each SNP functional class; see also Table 3). Non-CNEs displayed on the contrary a higher proportion of common SNPs (DAF>10%) (~52% versus 43% within CNEs) independent of their functional class (Figure 2; Table 3). Therefore, the low proportion of derived alleles in CNEs indicates that evolutionary pressure has suppressed CNE-derived allele frequencies.

Table 3: Derived allele frequency distribution for SNPs in CNEs and non-CNEs by SNP functional class.

Genomic feature	DAF	Within CNEs	Outside CNEs	chCADD within CNEs	chCADD outside CNEs
		Number of SNPs (%)	Number of SNPs (%)	Average (± sd)	Average (± sd)
All	≤0.10	137,871 (57%)	482,685 (48.4%)	9.78 (4.18)	3.21 (3.18)
	>0.10	103,726 (43%)	513,935 (51.5%)	8.81 (4.25)	2.74 (2.83)
LncRNA	≤0.10	24,364 (57.4%)	26,429 (47.6%)	10.02 (4.00)	3.49 (3.33)
	>0.10	18,081 (42.5%)	29,014 (52.4%)	9.10 (4.13)	3.03 (2.99)
Intron	≤0.10	43,790 (56.8%)	159,203 (47.4%)	9.81 (4.46)	3.00 (3.11)
	>0.10	33,171 (43.2%)	176,650 (52.6%)	8.71 (4.53)	2.46 (2.74)
Intergenic	≤0.10	69,717 (57%)	297,053 (44.6%)	9.68 (4.05)	3.31 (3.20)
	>0.10	52,474 (43%)	308,271 (55,4%)	8.78 (4.11)	2.87 (2.86)

5.3.4. chCADD scores for the investigation of CNE and SNP evaluation

To investigate CNEs further, we developed a model that can evaluate individual SNPs or entire sequences based on a per-base score, with respect to its putative deleteriousness. This model is based on the CADD approach, hence it is labeled ch(icken) CADD. chCADD is a linear logistic model that is trained to differentiate between two classes of variants, one being relatively more enriched in potentially deleterious variants than the other. To obtain these two classes, one class is generated from derived variants, alleles that have accumulated since the last ancestor with turkey and became fixed or almost fixed (>90% AF) in our chicken populations. These are depleted in deleterious variants and can be assumed to be benign or at least neutral in their nature. The set of putative deleterious variants contains simulated de novo variants that are not depleted of deleterious variants. The feature weights obtained during training are shown in Supplementary file 2. Performance on a held out test set to determine an optimal penalization term are shown in Figure S5.

5.3.5. chCADD scores potentially causal variants higher

We evaluated the performance and applicability of chCADD on two different sets of variants before we annotated non-coding SNPs.

First, we assigned a chCADD score to all SNPs found in the genomes of the 169 chickens previously used in the SNP density and DAF analysis and compared these to functional predictions as annotated by the Ensembl VEP (Figure S6). To this end, we categorized VEP predictions into 14 categories (Table S1). The purpose of this was to test whether chCADD correctly scores SNPs with respect to their potential to cause a deleterious or phenotype-changing effect, as indicated (mostly for protein-coding mutations) by the VEP functional predictions. We observed that mutations with a relatively large deleterious potential, such as stop-gained mutations and splice-site altering mutations, were scored higher than regular missense and synonymous mutations (Figure S6). SNPs in potentially regulatory active regions were also evaluated to be potentially more deleterious than synonymous SNPs (Figure S6). We performed a similar analysis considering only protein-coding and regulatory mutations found in the Online Mendelian Inheritance in Animals (OMIA) database [37] (Table 4). We annotated only SNPs whose genomic positions were uniquely mapped to the chicken GRCg6a reference genome and the reference/alternative allele matched that in the genome assembly. Of the 15 annotated SNPs associated with a change of phenotype, 5 were reported to cause a deleterious phenotype change in the affected individual, and an average chCADD score of 27.1. These 5 variants (3 stop-gained, 2 missense) have a chCADD score above 20 and are putatively responsible for dwarfism, scaleless, analphalipoproteinaemia, muscular dystrophy, and wingless phenotypes (Table 4). All these phenotypes display a strong severity and may lead to an early death in uncontrolled environments.

Table 4: OMIA chicken SNPs with chCADD annotations, locations are reported for Gal6.

OMIA ID(s)	Variant Phenotype	Gene	Type of Variant	Deleterious?	g. or m.	chCADD
OMIA 001622-9031	Resistance to avian sarcoma and leukosis viruses, subgroup C	BTN1A1	stop-gain	no	28:g.903289G>T	17.83409
OMIA 000889-9031	Scaleless	FGF20	stop-gain	yes	4:g.63270401A>T	33.02083
OMIA 001534-9031	Resistance to myxovirus	MX1	missense	no	1:g.110260061G>A	14.26893
OMIA 000915-9031	Feather colour, silver	SLC45A2	missense	no	Z:g.10336596G>T	21.72641
OMIA 000915-9031	Feather colour, silver	SLC45A2	missense	no	Z:g.10340909T>C	15.69336
OMIA 000679-9031	Muscular dystrophy	WWP1	missense	yes	2:g.123014353G>A	26.29866
OMIA 000303-9031	Dwarfism, autosomal	C1H12ORF23	stop-gain	yes	1:g.53638233C>T	35.29646
OMIA 001302-9031	Resistance to avian sarcoma and leukosis viruses, subgroup B	TNFRSF10B	stop-gain	no	22:g.1418711C>T	17.63145
OMIA 000810-9031	Polydactyly	LMBR1	regulatory	yes	2:g.8553470G>T	17.41378
OMIA 000913-9031	Silky/Silkie feathering	PDSS2	regulatory	unknown	3:g.67850419C>G	3.8812
OMIA 001547-9031	Wingless-2	RAF1	stop-gain	yes	12:g.5374854G>A	23.44641
OMIA 000374-9031	Feather colour, extended black	MC1R	missense	no	11:g.18840857T>C	18.05882
OMIA 000374-9031	Feather colour, extended black	MC1R	missense	no	11:g.18840919G>A	18.88983
OMIA 000374-9031	Feather colour, buttercup	MC1R	missense	no	11:g.18841289A>C	17.41773
OMIA 000374-9031	Feather colour, extended black	MC1R	regulatory; 5 'UTR	no	11:g.18840609C>T	6.74322

5.3.6. chCADD detects evolutionary constraints within CNEs

As we showed, chCADD can score functionally important protein-coding variants. We therefore decided to take a step further by annotating SNPs found in CNEs with chCADD to predict their deleteriousness and function (Table 3). We assume that highly scored SNPs can help us to identify truly functionally active regions among CNEs. We observed that rare non-protein-coding variants located within CNEs (DAF ≤ 10%) have an overall higher chCADD score compared to rare variants found in non-CNEs (Table 3). This result supports our previous conclusion based on the derived allele frequency spectrum that evolutionarily conserved non-protein-coding variants are likely

functional. As expected, this trend was most pronounced in lncRNAs, followed by introns and intergenic regions.

We further used the chCADD score to identify specific subregions of potentially higher functional importance within each CNE, assuming that the high scoring SNPs would indicate that. We applied a change point analysis to search for a center region that has high chCADD scores as opposed to the two outer regions (see Methods). We ranked CNEs based on positive chCADD score differences between the center region and the outer regions and filtered for significant difference (p-value of ≤ 0.05, t-test). The top 3 ranked CNEs that overlap with lncRNAs, intronic and intergenic regions, respectively, are shown in Figure 3A.1, B.1 and C.1.

Analogous to this subregion analysis based on chCADD score, we performed a subregion analysis based on the 23 sauropsids PhastCons scores. A.2-C.2 show the identified regions for the PhastCons score for the same CNEs as Figure 3A.1, 4C.1, respectively. These figures indicate that chCADD generates more discriminative subregions than PhastCons. Particularly interesting are the chCADD scores for the top intergenic regions (C.1). The chCADD score increased from ~5 to ~15 at the subregion change point. This is equal to an increase of predicted deleteriousness by one magnitude, from the top 33% highest scored sites in the entire genome to the top 3%.

To further investigate the subregion partitioning of the CNEs, we computed the SNP density in each region, for both the chCADD induced regions (Figure 4, blue bars) as well as the 23 sauropsids PhastCons induced regions (Figure 4, orange bars). In both bases, the SNP densities of the center region are lower than those of the outer regions. Moreover, all CNE subregions display a lower density than regions up- and downstream the CNE, supporting the functional importance of the CNEs in general. Interestingly, the center regions, as identified by the chCADD score, have in general a ~0.07% lower SNP density than the center regions detected using the PhastCons scores. Therefore, our findings suggest that chCADD is more effective in pinpointing potentially regions of interest.

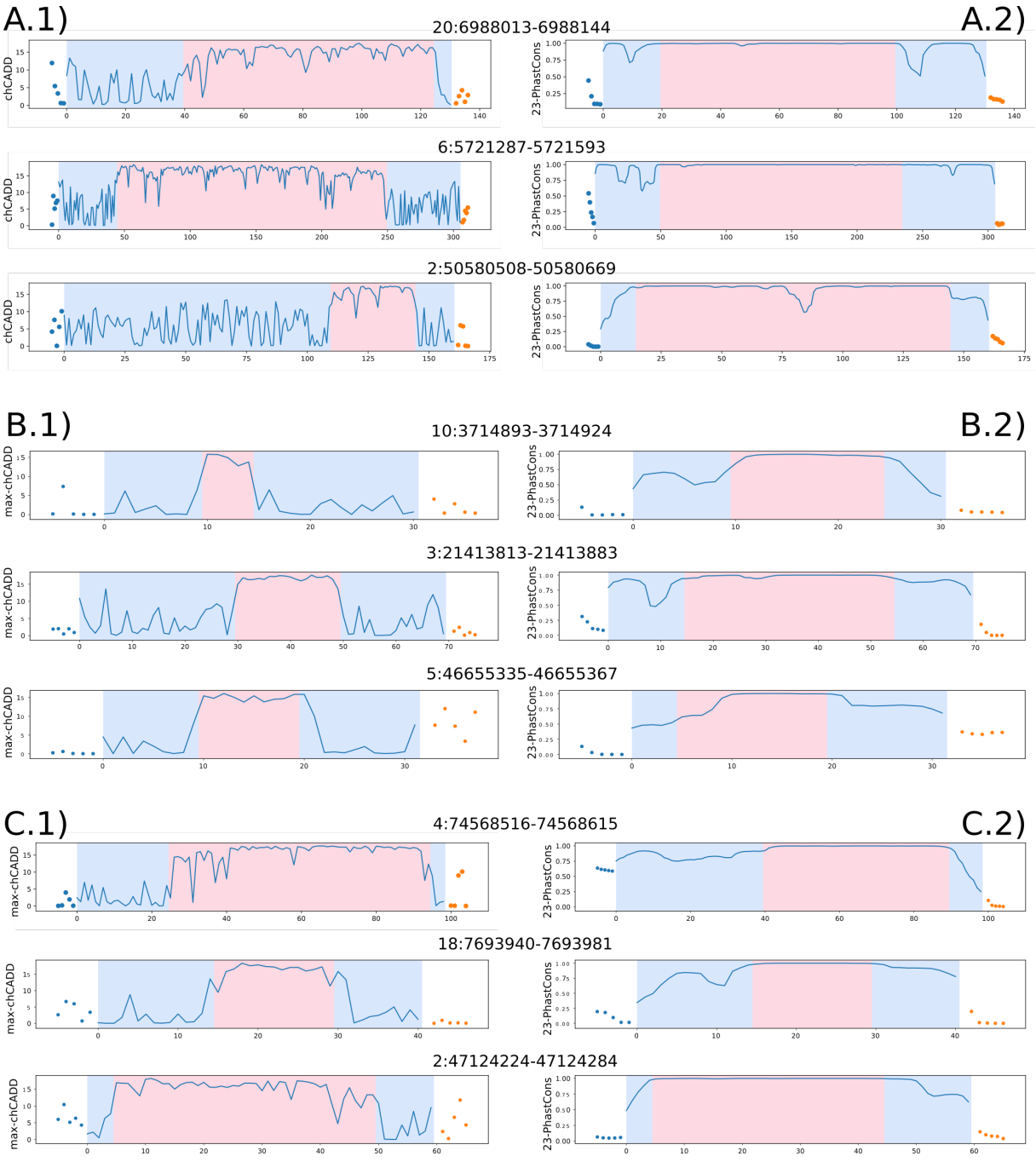


Figure 3: Change point analysis plots of the top 3 CNE regions for each CNE class respectively (lncRNA, intronic, intergenic). The CNE regions are sorted based on the largest difference between the 2nd section and 1st or 3rd section for each of the three CNE classes respectively (lncRNA, intronic, intergenic). The change points were once computed based on maximum chADD score per site (A.1,B.1,C.1) and once on 23 sauropsids PhastCons scores (A.2,B.2,C.2). The dots in each plot display the scores for the 5bp up- and downstream regions. The transition from blue to red background indicates the identified change points. A.1) lncRNA - maximum chCADD A.2) lncRNA - PhastCons scores. B.1) intronic - maximum chCADD. B.2) intronic - PhastCons. C.1) intergenic - maximum chCADD. C2) intergenic - PhastCons.

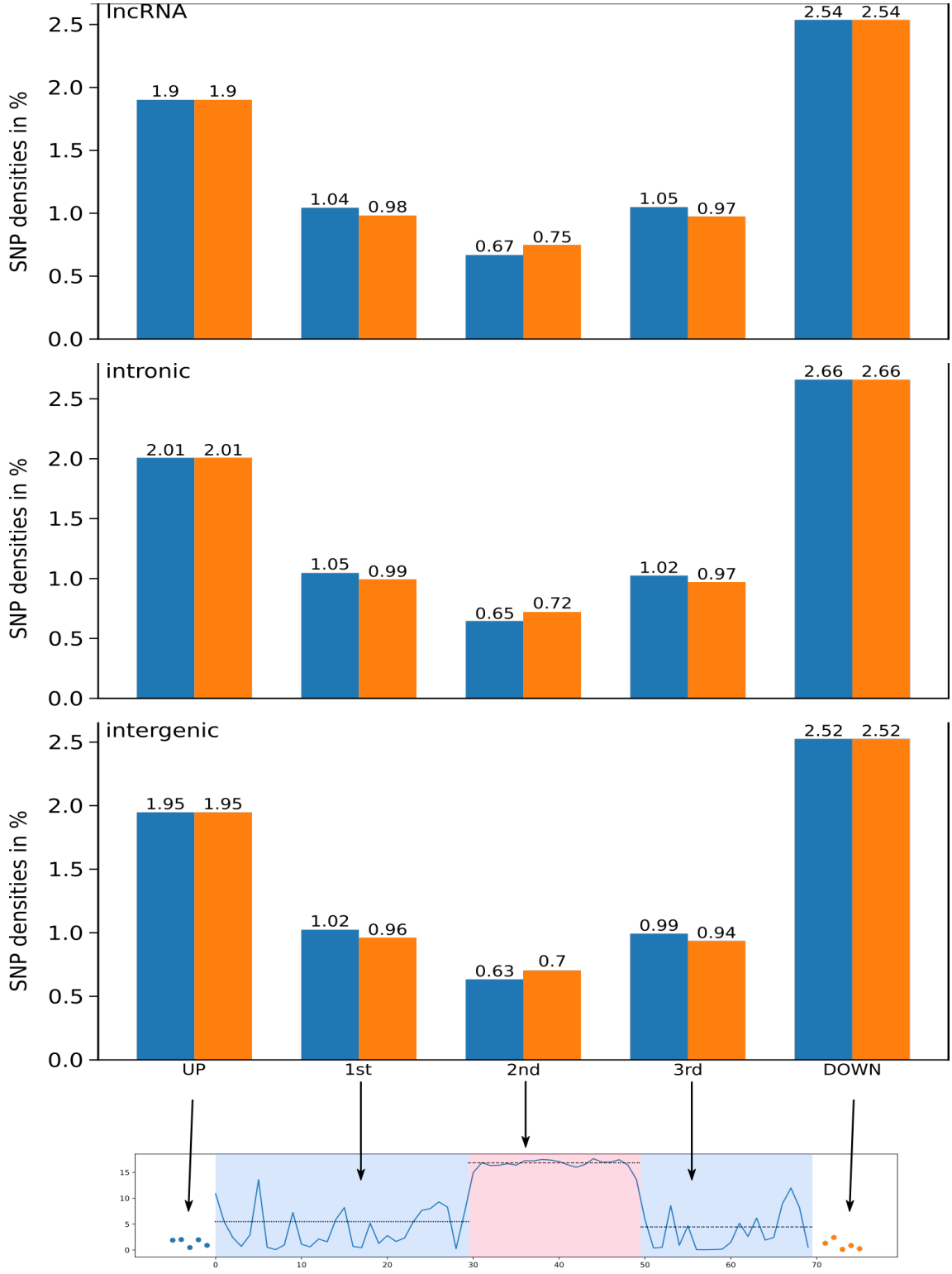


Figure 4: SNP densities computed for each section of the three different CNEs (lncRNA, Intronic, Intergenic). The orange bars represent the SNP densities for that section based on change points derived from 23 sauropsids alignment PhastCons scores, the blue bars represent the SNP densities based on change points identified via chCADD.

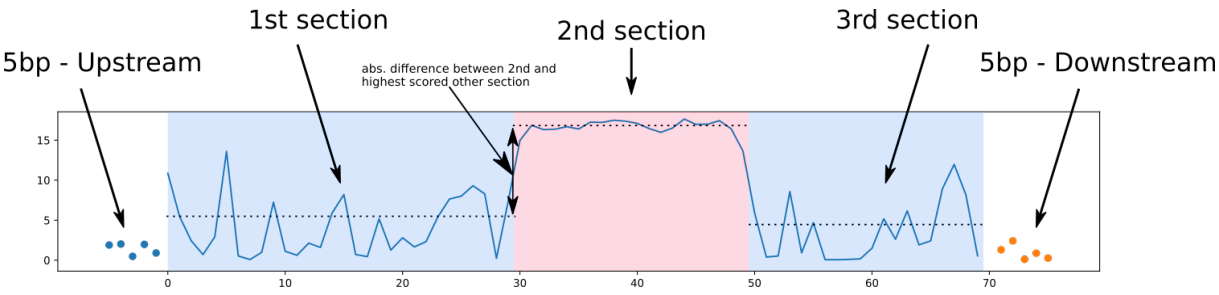


Figure 5: Approach used to identify subregions within CNEs via change point analysis. The scores used to annotate the CE region are displayed on the y-axis. The position in the investigated CE region is shown on the x-axis. In total there are five sections, 5 bp up and downstream, 1st, 2nd and 3rd subregions. The transitions from blue to red background indicate the position of the two identified change points. The up and downstream scores are shown as dots while the scores in the CE region are a continuous blue line.

5.3.7. Conserved non-protein-coding subregions are detected on the basis of a limited number of genomic annotations

As part of the investigation into subregions we identified two change points, splitting each CE into three subregions, starting from 5' to 3', 1st-, 2nd- and 3rd subregion (Figure 5). Next we were interested how genomic annotations that were used in the creation of chCADD, differ between the three subregions. The model coefficients with the largest weights (Table S2) point to the importance of the PhastCons conservation scores calculated on the 4 sauropsids alignment. Other important model features are secondary structure predictions and combinations with the intronic identifier from VEP. Over all CNEs, we compared the chCADD model features, especially the conservation scores that are based on different phylogenies, excluding the chicken reference sequence in their computation. For all genomic annotations, we computed absolute Cohen's D values (standardized mean difference) [38], [39]. We observed that the conservation scores based on the largest 77 vertebrate alignments cannot properly distinguish between the 1st-,2nd- and 3rd subregions. Conservation scores based on smaller phylogenies (4 sauropsids and 37 amniote/mammalia) are more discriminative between these (Table 5; see columns 1st-2nd, 2nd-3rd).

Considering the three PhastCons scores, based on differently large phylogenies, the average absolute Cohen's D between the 1st- and 2nd- and the 2nd- to the 3rd- subregions differ less between different genomic features (intergenic, lncRNA and introns) than between genomic annotations (Table 5; see columns 1st-2nd, 2nd-3rd). The average absolute Cohen's D between the three subregions of a CNE ranges from 0.259 to 0.276. In comparison, the average absolute Cohen's D between the same subregions, taking the three conservation scores individually, range from 0.137 to 0.338. The effect sizes between the different multiple sequence alignment PhastCons score (i.e. 4 sauropsids, 37 amniote/mammalia, 77 vertebrates) differ by more than 2-fold.

Table 5: Differences between genomic annotations utilized for the chCADD model, between CNE subregions defined by chCADD located in intronic, lncRNA and intergenic regions, measured in absolute Cohen's D.

INTRONIC	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	0.594	0.307	0.361	0.609
37PhastCons	0.446	0.328	0.369	0.448
77PhastCons	1.25	0.096	0.195	1.32
4PhyloP	0.43	0.09	0.126	0.428
37PhyloP	0.351	0.187	0.214	0.35
77PhyloP	0.776	0.186	0.237	0.778
GerpS	0.272	0.182	0.196	0.257
GerpN	0.212	0.112	0.11	0.214
dnaMGW	0.103	0.009	0.007	0.104
dnaProT	0.08	0.013	0.012	0.08
dnaHelT	0.082	0.002	0.002	0.083
GC	0.121	0.045	0.047	0.12
CpG	0.034	0.034	0.034	0.034
OChrom-Peaknb	0.058	0.001	0.091	0.015
OChrom-logFC	0.062	0.087	0.138	0.017
OChrom-pval	0.006	0.013	0.070	0.055
lncRNA	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	0.608	0.289	0.338	0.623
37PhastCons	0.469	0.31	0.342	0.482
77PhastCons	1.29	0.086	0.184	1.37
4PhyloP	0.428	0.083	0.117	0.43
37PhyloP	0.343	0.161	0.18	0.348
77PhyloP	0.788	0.17	0.22	0.792
GerpS	0.267	0.17	0.181	0.259
GerpN	0.212	0.086	0.098	0.201
dnaMGW	0.097	0.006	0.008	0.095
dnaProT	0.096	0.009	0.009	0.093
dnaHelT	0.089	0.003	0.0	0.086
GC	0.114	0.037	0.041	0.109
CpG	0.024	0.033	0.029	0.028
OChrom-Peaknb	0.059	-0.02	0.064	0.023
OChrom-logFC	0.102	0.093	0.137	0.055
OChrom-pval	0.012	0.096	0.103	0.005
INTERGENIC	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	0.61	0.281	0.341	0.619
37PhastCons	0.474	0.319	0.359	0.481
77PhastCons	1.29	0.084	0.179	1.37
4PhyloP	0.431	0.084	0.119	0.432

37PhyloP	0.351	0.162	0.185	0.351
77PhyloP	0.79	0.167	0.215	0.795
GerpS	0.29	0.169	0.183	0.274
GerpN	0.209	0.091	0.088	0.215
dnaMGW	0.096	0.008	0.008	0.096
dnaProT	0.097	0.014	0.012	0.096
dnaHelT	0.086	0.003	0.002	0.084
GC	0.136	0.062	0.062	0.136
CpG	0.039	0.037	0.036	0.041
OChrom-Peaknb	0.017	0.004	0.02	0.005
OChrom-logFC	0.089	0.005	0.012	0.077
OChrom-pval	0.00	0.005	0.052	0.023

5.3.8. Intronic CNE, differentially scored between the 1st , 2nd and 3rd subregions overlap functionally important genes

Intronic CNEs were associated with genes for which we obtained phenotype annotations of their orthologs in human, mouse, and rat. We investigated the top 10 CNEs that are located in introns, with the largest p-value differences between the 1st and 3rd to the 2nd section. 6 CNEs were associated with homologous genes that have annotated phenotypes in other species. Among the phenotypes found for human genes are mental retardation and non-syndromic male infertility. For mouse, these included neuronal issues and abnormal shape of heart and limbs (Table S3). The link to highly severe phenotypes in other species highlights the potential importance of regulatory features for orthologous genes in chicken.

5.4. Discussion

5.4.1. The prediction of CNEs depend on the phylogenetic scope

Non-protein-coding elements are typically identified by sequence-level similarity across species, which is a generally applicable criterion of conservation and biological function [10]. However, when predicting CEs, and subsequently CNEs, the evolutionary distance among species included in the alignment (or phylogenetic scope) is an important parameter that can considerably affect the prediction and resolution of CEs. If the evolutionary distance among species is too narrow, the specificity of constraint is reduced, but if it is too broad, the number of CEs rapidly declines and lineage-specific conservation is lost [10], [40].

One of the first studies to address the impact of the phylogenetic scope on CEs prediction was that of [12]. In their study on the 29 mammalian multiple sequence alignment the authors identified 3.6 million conserved elements spanning 4.2% of the genome at a resolution of 12 bp [12]. When comparing these results to a 5-vertebrate alignment, Lindblad-Toh and colleagues observed that only 45% of the 5-taxa CEs were covered by the 29-taxa alignment. This partial overlap indicates that most of the CEs derived from the 29-taxa alignment were mammalian-specific [12]. The issue resulting from a broad phylogenetic scope on CNEs has also recently been reported by [41], where authors identified CNEs between chicken and four mammalian species, including human, mouse, dog, and cattle [41]. By applying a minimum length of 100 bp, Babarinde and Saitou (2016) identified 21,584 CNEs in chicken, a small number as expected from the divergence time between

human and chicken ~310 million years ago [33]. Therefore, CNEs detected among distant species are better predictions of ultraconserved CNEs than CNEs between closely related species (i.e. human-mouse) [42], as they were already present in the ancient common ancestor of the considered species.

In this study we chose the 23 sauropsids multiple sequence alignment for two reasons. First, the phylogenetic distance between crocodilian and bird species (240 million years ago) [43] is large enough to detect likely functional CNEs. Second, the alignment is reference free allowing the identification of lineage-specific CEs. Reference-free alignments should always be preferred over reference-based ones [44]. In fact, genomic regions shared within a certain clade, which would be missed in a reference-based alignment (e.g. MULTIZ), can also be detected. As a result, reference free alignments better enable the study of genome evolution along all phylogenetic branches equally.

5.4.2. Avian genomes have similar genomic characteristics

According to our study, 8% of the chicken genome is covered by CEs for a total of 1.14 million CEs. These results are comparable to those on the collared flycatcher genome (Ficedula albicollis) [8]. By means of the same alignment, [8] identified 1.28 million CEs covering 7% of the flycatcher genome. Compared to the flycatcher, the slightly lower number of CEs we report in chicken could be explained by its smaller genome size, as small genomes require fewer regulatory sequences involved in the organization of chromatin structure [8]. For instance, the chicken genome is nearly 4 times smaller (i.e. GRCg6a: 1.13 Gb) than that of human (i.e. GRCh38.p13: 4.53 Gb), but of nearly equal size to that of the collared flycatcher (i.e. FicAlb1.5: 1.11 Gb). The similarity in genome size between chicken and flycatcher reflects the little cross-species variation characteristic of birds [45].

The limited number of CEs often identified in birds relative to mammals has repeatedly been linked to gene loss [23], [25]. However, the role of gene loss in avian evolution, genome size, and prediction of CEs has recently been questioned. According to [26], gene loss was incorrectly hypothesized from the absence of genes clustering in GC-rich regions in the earlier chicken genome assemblies [26]. In fact, these regions are often difficult to sequence and assemble. This issue is particularly prominent in the GC-rich micro-chromosomes, which, as we show, contribute disproportionately to the total density of functional sequence (Figure S1). We therefore recommend future comparative genomics studies in chicken to make use of the most recent and complete genome assembly to avoid any erroneous link of CEs to gene loss in chicken.

5.4.3. Conserved non-protein-coding elements are maintained by purifying selection

A fundamental question in the study of CNEs is the role of purifying selection. Purifying selection can be discriminated from a low mutation rate by comparing the derived allele frequency (DAF) spectra in constrained regions (i.e. CNEs) with that of neutral regions (i.e. non-CNEs) (9,35). This is because new mutations are unlikely to increase in frequency in constrained regions. Although CNEs are identified using an interspecific comparative genomic approach, the evolution and dynamics of these regions are generally analyzed at an intraspecific scale by looking at polymorphism data [9], [46]. In this study, we showed that the evolutionary constraint acting on the 23 sauropsids is correlated with constraint within the chicken populations, as assessed from chicken polymorphism data. Consistent with studies in humans [12], [35], plants [6], and Drosophila [9], [36], the derived allele frequency spectra of our chicken populations is shifted towards an excess of rare variants in CNEs. These results indicate that the conservation of CNEs in

the chicken genome is mainly driven by selective constraints, and not by local variation in mutation rate. The role of purifying selection was also confirmed by the reduced SNP density in CNEs compared to non-CNEs and by the reduced SNP density in specific conserved non-protein-coding subregions. The concordance in SNP density is a clear indication of reduced levels of population diversity and functional roles of CNEs as confirmed by the association of subregions within CNEs to highly severe phenotypes in humans, mouse, and rat. However, future population diversity comparisons in terms of nucleotide diversity (π) [47] or Watterson's estimator (θw) [48] between outbred and inbred populations would further elucidate our understanding of purifying selection in CNEs.

5.4.4. Integrating comparative and functional genomics into a single score

We developed a ch(icken) Combined Annotation-Dependent Depletion (chCADD) approach that provides scores for all SNPs throughout the chicken genome. These scores are indicative of putative SNP deleteriousness and can be used to prioritize variants.

The annotation of chCADD relies on the combination of a diverse set of genomic features, including evolutionary constraints and functional data [21], [22]. Multiple sequence alignments of distantly related species are better suited to differentiate conserved sites that can reliably be used to identify functionally important regions. However, these regions are often large enough to question the functional role of the entire region. Our findings show that chCADD outperforms any conservation-based method alone (e.g. PhastCons) in the identification of functionally important subregions within CNEs. Therefore, methods, such as chCADD, are required to fine-tune in one step CNEs to identify subregions directly linked to - in some cases deleterious – phenotypes.

According to the authors of the original human CADD [21], SNPs with a score above 20 (i.e. the SNP is among the top 1% highest scored potential SNPs in the genome) could be considered deleterious. This means that the higher the score, the higher the chance the variant has a functional effect or may even be deleterious. When annotating protein-coding and regulatory mutations found in OMIA [37], we observed that SNPs with a chCADD score of 15 can already be considered functional. Therefore, our findings indicate that by setting an arbitrary threshold of 20 may underestimate the fraction of the genome that is actually functional. This is particularly pronounced when the variants in question are located outside protein-coding regions. Therefore, we recommend future chCADD users to evaluate the variants identified in their populations to see if they are particularly highly scored compared to other variants in the same genomic region.

5.4.5. Future uses of chCADD

The high scoring of non-protein-coding variants in subregions of CNEs has important implications for future functional and genome-wide association studies (GWAS) in chicken. A very large fraction of trait- or disease-associated loci identified in GWAS are intronic or intergenic. This is expected considering the preponderance of non-protein-coding SNPs on genotyping arrays [5] or along the genome. However, because of a lack of understanding of the function of non-protein-coding mutations, most of the causal mutations reported in the OMIA database are coding. Moreover, in the presence of non-protein-coding mutations, many studies stop at the general locus or - understandably - assume that the closest neighbouring gene is affected. However, these assumptions on genomic distance are simplistic. Our findings in chicken demonstrate that chCADD can accurately pinpoint non-protein and protein-coding variants associated with important phenotypes in chicken. Therefore, we expect future genome-wide association studies combined with chCADD to identify novel causal mutations or substantially narrow down the list of potential

causal variants in large quantitative trait loci (QTLs). We also expect chCADD to accelerate the discovery and understanding of the biology and genetic basis of phenotypes.

5.5. Conclusions

Deciphering the function of the non-coding portion of a species genome has been a challenging task. However, the availability of genomes from a great variety of species, along with the development of new computational approaches at the interface of machine learning and bioinformatics, has made this task possible in model and non-model organisms. Our findings indicate an accurate assessment of selective pressure at individual sites becomes an achievable goal. We have also shown that chCADD is a reliable score for the analysis of non-protein-coding SNPs, which should be targeted along with protein-coding mutations in future genome-wide association studies. We therefore anticipate chCADD to be of great use to the scientific community and breeding companies in future functional studies in chicken.

5.6. Methods

5.6.1. Chicken genomic data

We used a dataset by Bortoluzzi and colleagues available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession number PRJEB34245 [49] and PRJEB36674 [18]. The 169 chicken samples included in the dataset were sequenced at the French Institute of Agricultural Research (INRA), France, on an Illumina HiSeq 3000. Reads were processed following standard bioinformatics pipelines. Reads were aligned to the chicken GRCg6a reference genome (GenBank Accession: GCA_000002315.5) with the Burrows-Wheeler alignment (BWA-mem) algorithm v0.7.17 [50]. After removal of duplicate reads with the markdup option in sambamba v0.6.3 [51], we performed population-based variant calling in Freebayes [52], retaining only sites with a mapping and base quality >20. We reduced the false discovery rate by additional filtering using BCFtools v1.4.1 [50].

5.6.1.1. Multiple whole-genome sequence alignment

Conserved elements (CE) were identified using the 23 sauropsids multiple whole-genome sequence alignment (MSA) generated using Progressive Cactus (https://github.com/glennhickey/progressiveCactus) [53] by [43]. The MSA downloaded in the hierarchical alignment format (HAL) was converted into multiple alignment format (MAF) using the HAL tools command hal2maf [54] with the following parameters: -refGenome galGal4 (GenBank Accession: GCA_000002315.2) to extract alignments referenced to the chicken genome assembly, -noAncestors to exclude any ancestral sequence reconstruction, -onlyOrthologs to include only sequences orthologous to chicken, and -noDupes to ignore paralogy edges. During reformatting, only blocks of sequences where chicken aligned to at least two other species were considered for a total chicken genome alignability of 90.88%. Genomic coordinates were converted to the GRCg6a genome assembly using the pyliftover library in python v3.6.3.

5.6.1.2. Prediction of evolutionarily conserved elements

Conserved elements were predicted from the whole-genome alignment using PhastCons [55]. We chose PhastCons because this approach does not use a fixed-size window approach, but can take advantage of the fact that most functional regions involve several consecutive sites [56]. We first generated a neutral evolutionary model from the 114,709 four-fold degenerate (4D) sites previously extracted from the alignment by [43]. The topology of the phylogeny was also identical

to that derived by [43]. PhastCons was run using the set of parameters used by the UCSC genome browser to produce the 'most conserved' tracks (top 5% of the conserved genome): expected length = 45, target coverage = 0.3, and rho = 0.31 [32]. Conserved elements were subsequently excluded if falling or overlapping assembly gaps and/or if their size was < 4 bp.

5.6.1.3. Annotation of conserved elements by genomic feature

We use the Ensembl (release 95) chicken genome annotation files to extract sequence coordinates of CDS, exons, 5' and 3' UTRs, pseudogenes, and lncRNAs. Sequence information was extracted from 14,828 genes (out of the 15,636 genes found in the Ensembl annotation), as transcripts of these genes had a properly annotated start and stop codon. For protein-coding genes with an annotated 5' UTR of at least 15 bp, the promoter was defined as the 2-kb region upstream of the transcription start site (TSS) [8]. Sequence coordinates of miRNAs, rRNAs, snoRNAs, snRNAs, ncRNAs, tRNAs, and scRNAs were also extracted from the annotation file. For the identification of intergenic regions, we considered all annotated protein-coding genes and defined intergenic regions as DNA regions located between genes that did not overlap any protein-coding genes in either of the DNA strands. The intersection between CEs and the various annotated genomic features was found following the approach of [12] of assigning a CE overlapping two or more genomic features to a single one in a hierarchical format: CDS, 5' UTR, 3' UTR, promoter, RNA genes, lncRNA, intronic, and intergenic region. Conserved non-protein-coding elements (CNEs) were defined as CEs without any overlap with exon-associated features (CDS, 5' UTR, 3' UTR, promoter, and RNA genes) and include lncRNAs, introns, and intergenic regions.

5.6.1.4. Gene ontology analysis

Genes in conserved regions overlapping CDS, 5' UTR, 3' UTR, and introns were separately used to perform a Gene Ontology analysis in g:Profiler [57] using Gallus gallus as organism. We only considered annotated genes that passed Bonferroni correction for multiple testing with a threshold < 0.05.

5.6.2. Genome-wide distribution and density of conserved non-protein-coding regions

CNE density and the density of exon-associated features were calculated in non-overlapping 100 kb windows along the genome. Windows that included assembly gaps between scaffolds were discarded, resulting in a total of 9,196 windows. Correlation between density of exons and CNEs was calculated in R v3.2.0 using the Pearson's correlation test.

5.6.3. Annotation of variants by functional class

Polymorphic, bi-allelic SNPs belonging to all functional classes predicted by the Variant Effect Predictor (VEP) [58] were considered. However, to improve the reliability of the set of annotated variants, we applied additional filtering steps. SNPs were discarded if they overlapped repetitive elements or if their call rate was <70%. The rationale for excluding variants found in repetitive elements was to reduce erroneous functional prediction as a result of mapping issues, as regions enriched for repetitive elements are usually difficult to assemble. Intronic and intergenic SNPs were further discarded if they overlapped spliced intronic ESTs [35]. Protein-coding variants were also discarded if they were found outside coding sequences, whose genomic coordinates were obtained from the Ensembl chicken GTF file (release 95).

5.6.4. Ancestral allele and derived allele frequency

The sequence of the inferred ancestor between chicken and turkey (Meleagris gallopavo; Turkey_2.01) [59] reconstructed from the Ensembl EPO 4 sauropsids alignment (release 95) was used to determine the ancestral and derived state of an allele, along with its derived allele frequency. We considered only SNPs for which either the reference or alternative allele matched the ancestral allele. Ancestral alleles that did not match either chicken allele were discarded. We generated derived allele frequency (DAF) distributions for sets of SNPs based on functional class and whether they were within or outside of CNEs. A derived allele frequency cutoff of 10% was used to distinguish rare from common SNPs.

5.6.5. Chicken Combined Annotation Dependent Depletion (chCADD)

The chicken CADD scores are the -10 log relative ranks of all possible alternative alleles of all autosomes and Z chromosome of the chicken GRCg6a reference genome, according to the following formula:

$$chCADD_i = -10\log_{10}\left(\frac{n_i}{N}\right)$$

where N represents the number of all possible alternative alleles (3,073,805,640) on the investigated chromosomes and n is the rank of the ith SNP. The ranks are based on the model posteriors of a ridge penalized logistic regression model trained to classify simulated and derived SNPs.

Chicken derived SNPs were defined as those sites where the chicken reference genome differs from the chicken-turkey ancestral genome inferred from the Ensembl EPO 4 sauropsids alignment. Sites for which the ancestral allele occurs at a minor allele frequency greater than 5% were excluded. In addition, derived SNPs that are observed with frequency above 90% in our population of 169 individuals were included. In total we identified 17,237,778 SNPs.

The dataset of simulated variants was simulated based on derived nucleotide substitution rates between the inferred ancestor of chicken, turkey, zebra finch (Taeniopygia guttata; taeGut3.2.4) [60] and green anole lizard (Anolis carolinensis; AnoCar2.0) [61]. These derived nucleotide substitution rates were obtained for windows of 100 kb and used to simulate de novo variants which have a larger probability to have a deleterious effect than the set of derived variants. All SNPs which have a known ancestral site are retained in the dataset. In total 17,233,727 SNPs were simulated in this way. 17,233,722 SNPs of each dataset were joined and randomly assigned to train and test sets of sizes 15,667,020 and 1,566,702, respectively.

The datasets were annotated with various genomic annotations: among others, PhyloP and PhastCons (Table S4) conservation scores based on three differently deep phylogenies (i.e. 4 sauropsids, 37 amniote/mammalia, 77 vertebrate, all excluding the chicken genome), secondary DNA structure predictions (Table S4), Ensembl Consequence predictions, amino acid substitution scores such as Grantham (Table S4) and amino acid substitution deleterious scores such as SIFT (Table S4).

Annotations for which values were missing were imputed, categorical values were one hot-encoded [62]. In the one hot-encoding process, an annotation is a series of binary annotations, each indicating the presence of a specific category for a given variant. For scores that are by definition not available for certain parts of the genome, such as SIFT which is found only for missense mutations, columns indicating their availability were introduced.

Combinations of annotations were created of Ensembl Variant Effect Predictor consequences and other annotations, such as distance to transcription start site and conservation scores. The total number of all features used in training was 874. An extensive list of all annotations, combinations of annotations and their learned model weights is shown in Supplementary File 2. Finally, each feature column is scaled by its standard deviation. The logistic regression is trained via the Python Graphlab module. We selected a penalization term of 1, based on results on the test set (Figure S5).

5.6.6. Investigation of likely causal SNPs from the OMIA database

We downloaded the likely causal variants of phenotype changes from the Online Mendelian Inheritance in Animals (OMIA) [37] database (last accessed 25.11.2019). SNPs whose location was reported for older genome assemblies such as Galgal4 and Galgal5 were mapped to the chicken GRCg6a reference genome via CrossMap [63]. We only consider bi-allelic SNPs whose genomic position was successfully mapped to GRCg6a and whose substitution remained the same. In total, 15 SNPs were left and annotated with chCADD.

5.6.7. Change point analysis

To identify sub-regions of particular importance within each CE, we annotated all with the maximum chCADD score found at each site or the 23-sauropsids PhastCons scores that were used to identify conserved elements in the first place. Our basic assumption was that highly important subregions within a CE are preceded and succeeded by less important sites which would result in a relatively higher score region surrounded by two lower scored regions. Each CE was treated similarly to time series data by conducting an offline change point analysis, once based on maximum chCADD scores and once based on 23-sauropsids PhastCons scores. To this end, we used the Python ruptures module [64] and applied a binary segmentation algorithm with radial basis function (RBF). It first identifies a single change point, if one is detected, the the algorithm investigates each sub-sequence independently to identify the next change point We were looking particularly for 2 change points, which would divide the CE into three subregions, numbered from 1 to 3, starting at the 5’ end of the sequence. We added 5 bp upstream and downstream of each CE to allow that the borders of the 2nd region coincide with the borders of the CE (Figure 5). After computing the change points, we conducted t-tests between the scores of the 1st and 2nd, as well as 3rd and 2nd subregions, to identify CEs that have a significantly different score in the 2nd section than in the other two. We applied a p-value cutoff of 0.05. We sorted CNEs with respect to the largest difference between the mean chCADD score of the inner and the two outer subregions and selected those with a higher scored 2nd section than either of the other two outer ones.

5.6.8. SNP density distribution within conserved non-protein-coding regions

SNP density was calculated as the number of SNPs identified in the 169 chicken individuals divided by the number of bases found in the sequence. SNP density was computed for conserved coding (CC) and conserved non-protein-coding (CNE) regions, as well as for the subregions identified in the change point analysis of CNEs overlapping lncRNAs, introns, and intergenic regions. We repeated this analysis once for the change points identified using chCADD scores and once for the 23-sauropsids PhastCons based change points.

5.6.9. Homologous phenotypes

We obtained phenotypes from the Ensembl database (release 95) for genes associated with the lncRNA and intronic CNEs. Beside chicken, these phenotypes encompass the observed phenotypes for orthologous genes associated with disease studies in humans (Homo sapiens) and gene-knockout studies in mouse (Mus musculus) and rat (Rattus norvegicus).

5.7. Declarations

5.7.1. Data access

Raw sequences the 169 individuals used in this study are available at the European Nucleotide Archive under accession number PRJEB34245 and PRJEB36674. chCADD scores (~GB) can be downloaded from the Open Science Framework project page (https://osf.io/8gdk9/).

5.7.2. Funding

C.G. was funded by the TTW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality. C.B was funded by the European Union´s Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353 (Innovative Management of Animal Genetic Resources – IMAGE). M.B. is financially supported by the NWO-VENI grant no.016.Veni.181.050.

Bibliography

[1] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] ENCODE Project Consortium, "The ENCODE (ENCyclopedia of DNA elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[3] ENCODE Project Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, no. 7146, p. 799, 2007.

[4] E. H. Margulies *et al.*, "Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome," *Genome Res.*, vol. 17, no. 6, pp. 760–774, 2007.

[5] R. P. Alexander, G. Fang, J. Rozowsky, M. Snyder, and M. B. Gerstein, "Annotating non-coding regions of the genome," *Nat. Rev. Genet.*, vol. 11, no. 8, pp. 559–571, 2010.

[6] A. Haudry *et al.*, "An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions," *Nat. Genet.*, vol. 45, no. 8, pp. 891–898, 2013.

[7] J. Alföldi and K. Lindblad-Toh, "Comparative genomics as a tool to understand evolution and disease," *Genome Res.*, vol. 23, no. 7, pp. 1063–1068, 2013.

[8] R. J. Craig, A. Suh, M. Wang, and H. Ellegren, "Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (Ficedula albicollis)," *Mol. Ecol.*, vol. 27, no. 2, pp. 476–492, 2018.

[9] T. Berr, A. Peticca, and A. Haudry, "Evidence for purifying selection on conserved noncoding elements in the genome of Drosophila melanogaster," *bioRxiv*, p. 623744, 2019.

[10] N. Harmston, A. Barešić, and B. Lenhard, "The mystery of extreme non-coding conservation," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 368, no. 1632, 2013.

[11] I. Braasch *et al.*, "The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons," *Nat. Genet.*, vol. 48, no. 4, pp. 427–437, 2016.

[12] K. Lindblad-Toh *et al.*, "A high-resolution map of human evolutionary constraint using 29 mammals," *Nature*, vol. 478, no. 7370, pp. 476–482, 2011.

[13] D. L. Halligan *et al.*, "Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents," *PLoS Genet.*, vol. 9, no. 12, 2013.

[14] R. D. Hernandez *et al.*, "Classic selective sweeps were rare in recent human evolution," *Science*, vol. 331, no. 6019, pp. 920–924, 2011.

[15] R. J. Williamson *et al.*, "Evidence for widespread positive and negative selection in coding and conserved noncoding regions of Capsella grandiflora," *PLoS Genet.*, vol. 10, no. 9, 2014.

[16] A. Marcovitz, R. Jia, and G. Bejerano, "'Reverse Genomics' predicts function of human conserved noncoding elements," *Mol. Biol. Evol.*, vol. 33, no. 5, pp. 1358–1369, 2016.

[17] M. T. Maurano *et al.*, "Systematic localization of common disease-associated variation in regulatory DNA," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[18] C. Bortoluzzi *et al.*, "Parallel genetic origin of foot feathering in birds," *Mol. Biol. Evol.*, vol. accepted, 2020.

[19] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, 2009.

[20] P. Park, "Applications of next-generation sequencing: ChIP–seq: advantages and challenges of a maturing technology," *Nat. Rev. Genet.*, vol. 10, no. 10, p. 669, 2009.

[21] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants.," *Nat. Genet.*, vol. 46, no. 3, pp. 310–5, 2014.

[22] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: Predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, 2019.

[23] G. Zhang *et al.*, "Comparative genomics reveals insights into avian genome evolution and adaptation," *Science*, vol. 346, no. 6215, pp. 1311–1320, 2014.

[24] R. W. Meredith, G. Zhang, M. T. P. Gilbert, E. D. Jarvis, and M. S. Springer, "Evidence for a single loss of mineralized teeth in the common avian ancestor," *Science*, vol. 346, no. 6215, 2014.

[25] P. V. Lovell *et al.*, "Conserved syntenic clusters of protein coding genes are missing in birds," *Genome Biol.*, vol. 15, no. 12, p. 565, 2014.

[26] S. Bornelöv *et al.*, "Correspondence on Lovell et al.: Identification of chicken genes previously assumed to be evolutionarily lost," *Genome Biol.*, vol. 18, no. 1, pp. 1–4, 2017.

[27] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Res.*, vol. 11, no. 5, pp. 863–874, 2001.

[28] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Curr. Protoc. Hum. Genet.*, vol. 2, 2013.

[29] Y. Choi and A. P. Chan, "PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics*, vol. 31, no. 16, pp. 2745–2747, 2015.

[30] C. Groß, D. de Ridder, and M. Reinders, "Predicting variant deleteriousness in non-human species: Applying the CADD approach in mouse," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, 2018.

[31] C. Groß *et al.*, "PCADD: SNV prioritisation in Sus scrofa," *Genet. Sel. Evol.*, vol. 52, no. 1, pp. 1–15, 2020.

[32] W. Miller *et al.*, "28-Way vertebrate alignment and conservation track in the UCSC Genome Browser," *Genome Res.*, vol. 17, no. 12, pp. 1797–1808, 2007.

[33] International Chicken Genome Sequencing Consortium, "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution," *Nature*, vol. 432, no. 7018, pp. 695–716, 2004.
[34] G. Bejerano *et al.*, "Ultraconserved elements in the human genome," *Science*, vol. 304, no. 5675, pp. 1321–1325, 2004.
[35] J. A. Drake *et al.*, "Conserved noncoding sequences are selectively constrained and not mutation cold spots," *Nat. Genet.*, vol. 38, no. 2, pp. 223–227, 2006.
[36] S. Casillas, A. Barbadilla, and C. M. Bergman, "Purifying selection maintains highly conserved noncoding sequences in Drosophila," *Mol. Biol. Evol.*, vol. 24, no. 10, pp. 2222–2234, 2007.
[37] J. Lenffer, "OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D599–D601, 2006.
[38] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Routledge, 1988.
[39] S. S. Sawilowsky, "New effect size rules of thumb," *J. Mod. Appl. Stat. Methods*, vol. 8, no. 2, pp. 597–599, 2009.
[40] G. M. Cooper and J. Shendure, "Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data," *Nat. Rev. Genet.*, vol. 12, no. 9, pp. 628–640, 2011.
[41] I. A. Babarinde and N. Saitou, "Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics," *Mol. Biol. Evol.*, vol. 33, no. 7, pp. 1807–1817, 2016.
[42] D. Polychronopoulos, J. W. D. King, A. J. Nash, G. Tan, and B. Lenhard, "Conserved non-coding elements: Developmental gene regulation meets genome organization," *Nucleic Acids Res.*, vol. 45, no. 22, pp. 12611–12624, 2017.
[43] R. E. Green *et al.*, "Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs," *Science*, vol. 346, no. 6215, 2014.
[44] J. Armstrong *et al.*, "Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era," *bioRxiv*, 2019.
[45] G. Zhang, "The bird's-eye view on chromosome evolution," *Genome Biol.*, vol. 19, no. 1, pp. 18–20, 2018.
[46] K. A. Steige, B. Laenen, J. Reimegård, D. G. Scofield, and T. Slotte, "Genomic analysis reveals major determinants of cis-regulatory variation in Capsella grandiflora," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 5, pp. 1087–1092, 2017.
[47] M. Nei and W. H. Li, "Mathematical model for studying genetic variation in terms of restriction endonucleases," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 76, no. 10, pp. 5269–5273, 1979.
[48] G. A. Watterson, "On the number of segregating sites in genetical models without recombination," *Theor. Popul. Biol.*, vol. 7, no. 2, pp. 256–276, 1975.
[49] C. Bortoluzzi, M. Bosse, M. F. L. Derks, R. P. M. A. Crooijmans, M. A. M. Groenen, and H. J. Megens, "The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations," *Evol. Appl.*, no. September 2019, pp. 330–341, 2019.
[50] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.
[51] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins, "Sambamba: Fast processing of NGS alignment formats," *Bioinformatics*, vol. 31, no. 12, pp. 2032–2034, 2015.
[52] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," pp. 1–9, 2012.
[53] B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, and D. Haussler, "Cactus: Algorithms for genome multiple sequence alignment," *Genome Res.*, vol. 21, no. 9, pp. 1512–1528, 2011.
[54] G. Hickey, B. Paten, D. Earl, D. Zerbino, and D. Haussler, "HAL: A hierarchical format for storing and analyzing multiple genome alignments," *Bioinformatics*, vol. 29, no. 10, pp. 1341–1342, 2013.
[55] A. Siepel *et al.*, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.," *Genome Res.*, vol. 15, no. 8, pp. 1034–50, 2005.
[56] J. Sadri, A. B. Diallo, and M. Blanchette, "Predicting site-specific human selective pressure using evolutionary signatures," *Bioinformatics*, vol. 27, no. 13, pp. 266–274, 2011.
[57] U. Raudvere *et al.*, "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W191–W198, 2019.
[58] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor.," *Genome Biol.*, vol. 17, no. 1, p. 122, 2016.
[59] R. A. Dalloul *et al.*, "Multi-platform next-generation sequencing of the domestic Turkey (Meleagris gallopavo): Genome assembly and analysis," *PLoS Biol.*, vol. 8, no. 9, 2010.
[60] W. C. Warren *et al.*, "The genome of a songbird," *Nature*, vol. 464, no. 7289, pp. 757–762, 2010.
[61] J. Alföldi *et al.*, "The genome of the green anole lizard and a comparative analysis with birds and mammals," *Nature*, vol. 477, no. 7366, pp. 587–591, 2011.
[62] H. Draper, N.R; Smith, *Applied regression analysis*. John Wiley & Sons, 1998.
[63] H. Zhao, Z. Sun, J. Wang, H. Huang, J. Kocher, and L. Wang, "CrossMap : a versatile tool for coordinate conversion between genome assemblies," vol. 30, no. 7, pp. 1006–1007, 2014.
[64] C. Truong, L. Oudre, and N. Vayatis, "Ruptures: change point detection in Python," pp. 1–5, 2018.

5.8. Appendix - Supplementary data

5.8.1. Supplementary figures

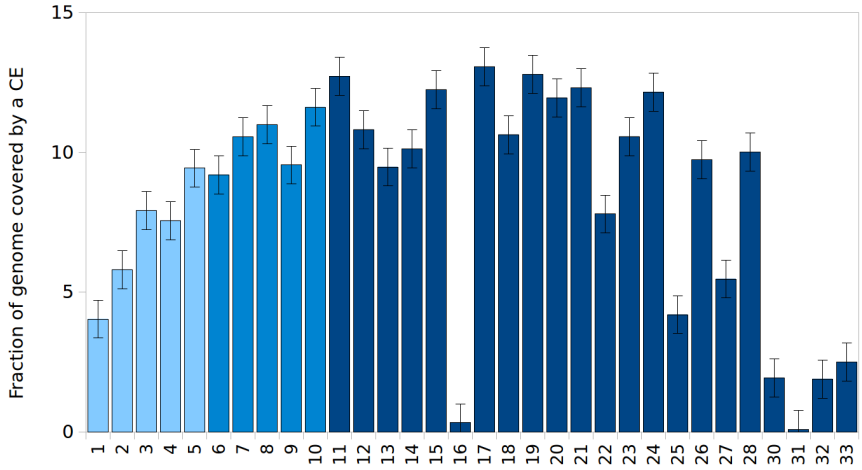


Figure S1: Distribution of conserved elements (CEs) along the chicken genome. The barplot displays the fraction of the genome per chromosome covered by conserved elements.

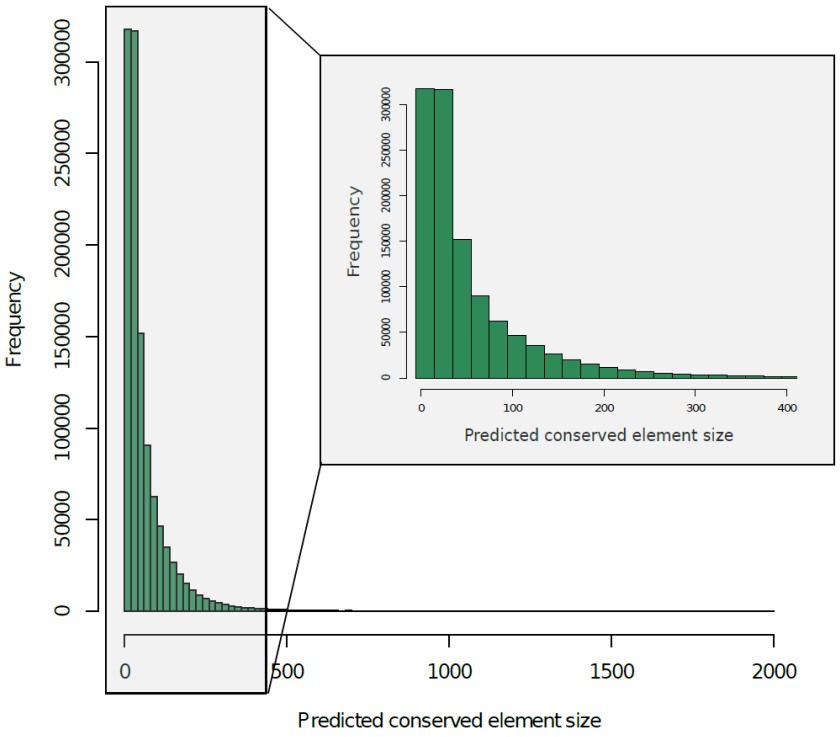


Figure S2: Frequency size distribution of predicted conserved elements. The y-axis shows the frequency, while the x-axis the size in base pairs (bp) of the predicted conserved elements.

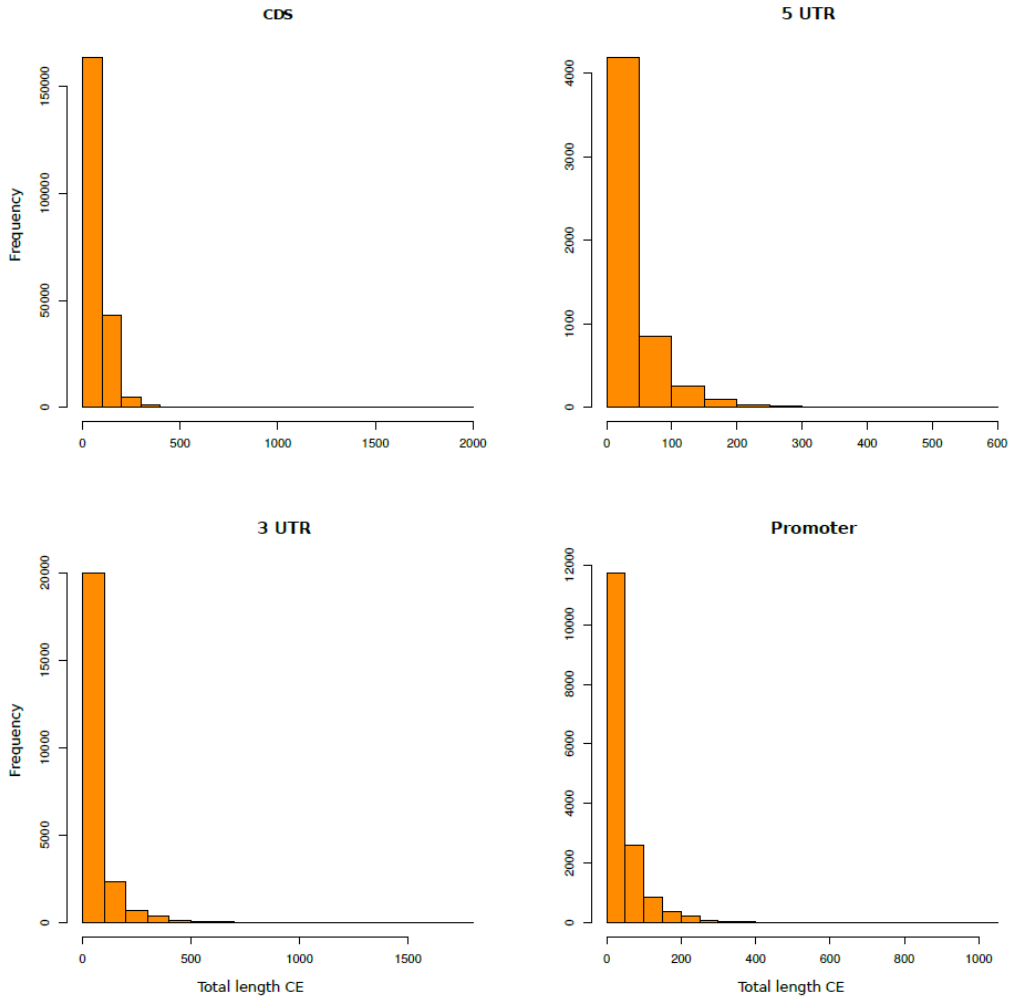


Figure S3: Frequency size distribution of predicted conserved elements overlapping exonic-associated gene annotations. The exonic-associated conserved elements include CDS, 5'UTR, 3'UTR, and promoter regions.

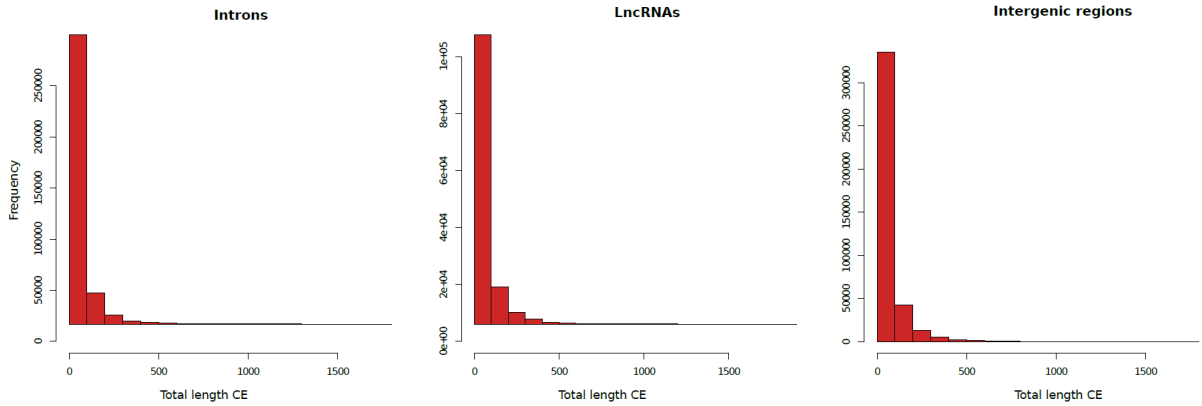


Figure S4: Frequency size distribution of predicted conserved elements overlapping non-protein-coding gene annotations. The non-protein-coding gene annotations include introns, lncRNA, and intergenic regions.

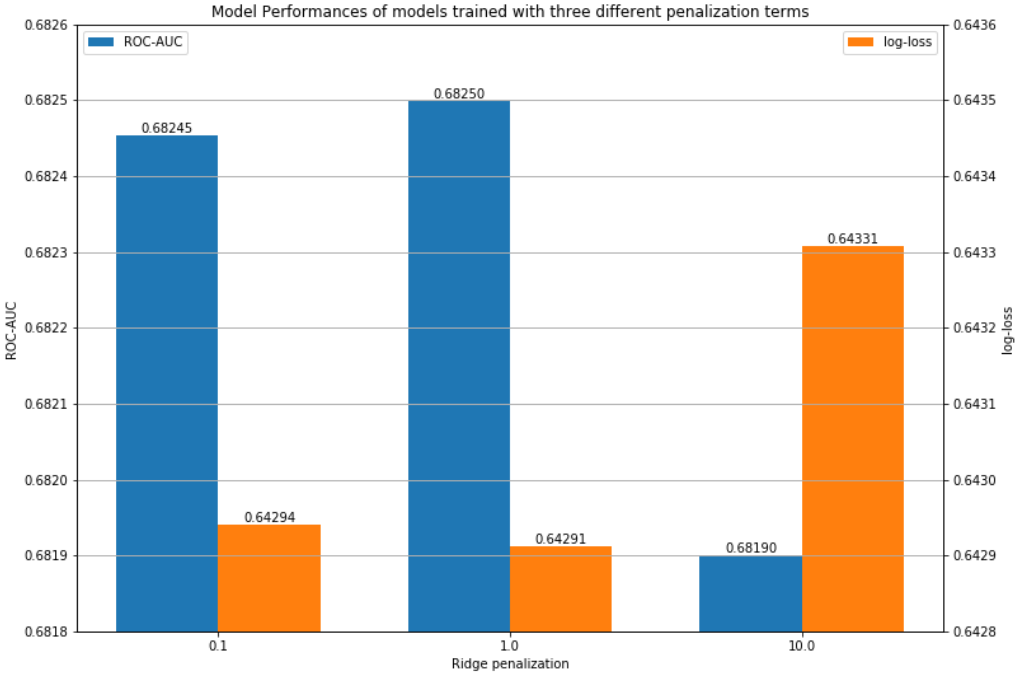


Figure S5: Model performances measured in Receiver Operator Area under the Curve (ROC-AUC) and log-loss for three different ridge penalization terms (0.1, 1.0, 10.0). The scale is adjusted to make the differences between the models visible. Penalization of 1 was.

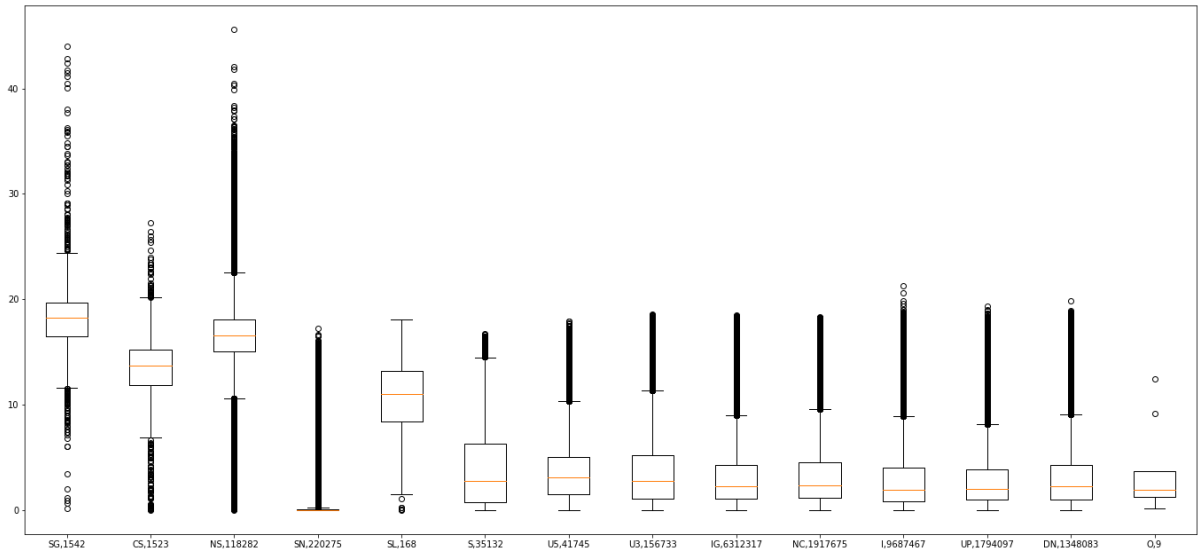


Figure S6: chCADD score distribution of SNPs per VEP cateogyr. SNPs from the 169 chickens are categorized based on the VEP categories reported in Table S1 (SG: Stop-gained; CS: Canonical Splice; NS: Non-Synonymous; SN: Synonymous; SL: STOP-Lost; S: Splice Site.

5.8.2. Supplementary tables

Table S1: VEP consequences summarized in 14 categories. If multiple annotations exist for the same variant, the consequence is selected according to the displayed hierarchy, starting at 1 and ending at 14.

Hierarchy	Abbreviation	VEP Consequence
1	SG	Stop Gained
2	CS	Canonical Splice
3	NS	Non-Synonymous
4	SN	Synonymous
5	SL	STOP Lost
6	S	Splice Site
7	U5	5'-UTR
8	U3	3'-UTR
9	IG	Intergenic
10	NC	Noncoding-change
11	I	Intronic
12	UP	Upstream
13	DN	Downstream
14	O	Unknown / Other

Table S2: Top 10 Model features with the largest assigned weight and their explanations.

Label	Model weight assigned to feature	Feature explanation
GerpS	0.152568	GERP rejected substitution score
4PhCons_noChick	0.128726	4-sauropsids PhastCons scores (excluding chicken)
I_GerpS	0.109099	GERP rejected substitution score for intronic sites
I_4PhCons_noChick	0.0899441	4-sauropsids PhastCons scores (excluding chicken) for intronic sites
dnaProT	0.083813	DNA secondary structure prediction for ProT
77PhCons_noChick	0.0790709	4-amniota PhastCons scores (excluding chicken
dnaRoll	0.0733429	DNA secondary structure prediction for Roll
IG_4PhCons_noChick	0.067539	4-sauropsids PhastCons scores (excluding chicken) for intergenic sites
I_dnaProT	0.0671401	DNA secondary structure prediction for ProT for intronic sites
IG_GerpS	0.0635293	GERP rejected substitution score for intergenic sites

Table S3: Phenotypes of homologous genes of the top 10 intronic CNEs. The top 10 intronic CNEs were selected based on the largest differences between the 1st and 3rd to the 2nd section within a CNE.

Chr	Start - End CE	Ensembl ID	Human phenotype	Mouse phenotype	Rat phenotype
10	3714893 - 3714924	ENSGALG00000002883	Autosomal Recessive Mental Retardation, intellectual developmental disorder and retinitis pigmentosa, Retinitis pigmentosa	abnormal heart left ventricle morphology, decreased grip strength, decreased large unstained cell number, decreased lean body mass, increased or absent threshold for auditory brainstem response, male infertility, preweaning lethality incomplete penetrance	-
3	21413813 - 21413883	ENSGALG00000009791	-	abnormal endocrine pancreas morphology, abnormal eye development, abnormal lens development, abnormal lens morphology, abnormal liver development, abnormal lymph organ development, absent horizontal cells, decreased hepatocyte proliferation, decreased lymphatic vessel endothelial cell number, edema, increased pancreatic acinar cell number, lethality throughout fetal growth and development complete penetrance, no abnormal phenotype detected, small liver, small pancreas	Status Epilepticus
5	46655335 - 46655367	ENSGALG00000011093	Non-syndromic male infertility due to sperm motility disorder, spermatogenic failure 27	abnormal cerebellum morphology, abnormal head shape, abnormal internal nares morphology, abnormal respiratory motile cilium morphology, abnormal respiratory motile cilium physiology, abnormal sperm head morphology, arrest of spermatogenesis, azoospermia, enlarged lateral ventricles, hydroencephaly, impaired mucociliary clearance, male infertility, oligozoospermia, postnatal growth retardation, premature death, respiratory system inflammation, rhinitis, thin cerebral cortex	-
5	29339796 - 29339843	ENSGALG00000009587	Hereditary hyperekplexia, hyperekplexia 1, Molybdenum cofactor deficiency complementation group C, Sulfite oxidase deficiency due to molybdenum cofactor deficiency type C	abnormal axon extension, abnormal motor neuron morphology, abnormal nervous system electrophysiology, abnormal neuromuscular synapse morphology, abnormal posture, abnormal retinal inner plexiform layer morphology, abnormal suckling behavior, abnormal vocalization, apnea, decreased motor neuron number, hyperresponsive, increased motor neuron number, motor neuron degeneration, neonatal lethality complete penetrance, no abnormal phenotype detected	inherited metabolic disorder
1	66913273 - 66913298	ENSGALG00000013244	Acromegaloid facial appearance syndrome, atrial fibrillation familial 12, brugada syndrome, cantu syndrome, cantu syndrome hypertrichotic osteochondrodysplasia, cardiomyopathy dilated 1o, familial atrial fibrillation, Familial isolated dilated cardiomyopathy, Hypertrichosis-acromegaloid facial appearance syndrome, Hypertrichotic osteochondrodysplasia Cantu type	abnormal ST segment, abnormal systemic arterial blood pressure, abnormal vascular smooth muscle physiology, abnormal vasoconstriction, artery stenosis, hypertension, hypoglycemia, improved glucose tolerance, increased insulin sensitivity, increased muscle cell glucose uptake, increased systemic arterial diastolic blood pressure, increased systemic arterial systolic blood pressure, premature death, slow postnatal weight gain	Diabetes Mellitus Experimental, hypertension, Parkinsonian Disorders, Sciatic Neuropathy, Ventricular Fibrillation, Ventricular Tachycardia

18	2616518 - 2616533	ENSGALG0000000137 5	isolated cytochrome c oxidase deficiency, leigh syndrome, mitochondrial complex iv deficiency	-	-
3	46314957 - 46314976	ENSGALG0000001225 6	-	abnormal miniature endplate potential, abnormal nervous system physiology, abnormal neuromuscular synapse morphology, increased sensitivity to xenobiotic induced morbidity/mortality	Duchenne muscular dystrophy, Ovarian Neoplasms
1	34590834 - 34590912	ENSGALG0000000989 5	Fraser syndrome, Fraser syndrome 3	abnormal blood coagulation, abnormal blood vessel morphology, abnormal brain morphology, abnormal corneal epithelium morphology, abnormal corneal stroma morphology, abnormal cornea thickness, abnormal eye development, abnormal eyelid morphology, abnormal eye morphology, abnormal iris morphology, abnormal kidney development, abnormal lens development, abnormal lens vesicle development, abnormal limb morphology, abnormal neural tube morphology, abnormal retina morphology, absent kidney, absent limbs, anophthalmia, aphakia, bleb, blistering, cataract, clubfoot, corneal opacity, decreased body size, embryonic lethality during organogenesis incomplete penetrance, eye hemorrhage, eyelids open at birth, hemorrhage, interdigital webbing, intracranial hemorrhage, kidney cysts, microphthalmia, open neural tube, perinatal lethality incomplete penetrance, polycystic kidney, polydactyly, prenatal lethality complete penetrance, single kidney, small kidney, syndactyly	-
2	102926228 - 102926256	ENSGALG0000001499 8	-	abnormal CNS glial cell morphology, abnormal endometrial gland morphology, abnormal endometrium morphology, absent corpus callosum, decreased litter size, dilated uterus, endometrium hyperplasia, enlarged uterus, increased endometrial carcinoma incidence, reduced female fertility	-
7	26190151 - 26190213	ENSGALG0000001164 5		preweaning lethality incomplete penetrance	

Table S4: List of annotations which form the set of descriptive features for which model weights are learned. Missing values are imputed via the specified values. Annotations of the type (factor) are OneHotEncoded and combinations between annotations form the final feature set.

Annotation label	Data type	Imputed value	Annotation description
Ref	factor		Reference allele
Alt	factor		Observed allele
isTv	bool	0.5	Is transversion?
Consequence	factor		VEP Consequence summaries
GC	num	0.4	Percent GC in a window of +/- 75bp
CpG	num	0.02	Percent CpG in a window of +/- 75bp
motifECount	int	0.0	Total number of overlapping motifs
motifEHIPos	bool	False	Is the position considered highly informative for an overlapping motif by VEP
motifEScoreChng	num	0.0	VEP score change for the overlapping motif site
Domain	factor	UD	Domain annotation inferred from VEP annotation (ncoils, tmhmm, sigp, lcompl, ndomain = "other named domain")
Dst2Splice	int	0.0	Distance to splice site in 20bp; positive: exonic, negative: intronic
Dst2SplType	factor	UD	Closest splice site is ACCEPTOR or DONOR
oAA	factor	UD	Amino acid of observed variant
nAA	factor	UD	Reference amino acid
Grantham	int	0.0	Grantham score: oAA,nAA
SIFTcat	factor	UD	SIFT category of change
SIFTval	num	0.0	SIFT score
cDNApos	int	0.0	Base position from transcription start
relcDNApos	num	0.0	Relative position in transcript
CDSpos	int	0.0	Base position from coding start
relCDSpos	num	0.0	Relative position in coding sequence
protPos	int	0.0	Amino acid position from coding start
relProtPos	num	0.0	Relative position in protein codon
dnaRoll	num	0.23	Predicted local DNA structure effect on dnaRoll
dnaProT	num	0.68	Predicted local DNA structure effect on dnaProT
dnaMGW	num	0.03	Predicted local DNA structure effect on dnaMGW
dnaHelT	num	-0.12	Predicted local DNA structure effect on dnaHelT
GerpS	num	-0.17	Rejected Substitution' score defined by GERP++
GerpN	num	0.64	Neutral evolution score defined by GERP++
GerpRS	num	0.0	Gerp element score
GerpRSpval	num	1.0	Gerp element p-Value
4PhCons_noChick	num	0.17	4-taxa-sauropsids PhastCons score (excl. chicken)
37PhCons_ noChick	num	0.13	37-taxa-Amniota PhastCons score (excl. chicken)
77PhCons_ noChick	num	0.2	77-taxa-Vertebrate PhastCons score (excl. chicken)
4PhyloP_ noChick	num	0.07	4-taxa-sauropsids PhyloP score (excl. chicken)
37PhyloP_ noChick	num	0.04	37-taxa-Amniota PhyloP score (excl. chicken)
77PhyloP_ noChick	num	0.25	77-taxa-Vertebrate PhyloP score (excl. chicken)
minDistTSS	int	10000000	Distance to closest Transcribed Sequence Start (TSS)
minDistTSE	int	10000000	Distance to closest Transcribed Sequence End (TSE)
interaction-score	num	0	Interaction score from Hi-C interaction maps
Exp-score	int	0	RNA expression scores
Exp-pval	num	1	p-Value of RNA expression scores
Exp-logFC	num	0	Log-Fold change of RNA expression
OChrom-Peaknb	Int	0	Read number for open Chromatin; ATAC-seq
OChrom-pval	num	1	p-Value for open chromatin; ATAC-seq
OChrom-logFC	num	0	Log-Fold change for ATAC-seq

6. Discussion

Variations in the genome can have a profound effect on the function or expression of genes. Due to many levels of interactions that can happen within a biological cell, it is infeasible to manually investigate all potentially important genomic regions that result from phenotype screens. Insights in the regulatory mechanisms of a cell, however, would greatly increase our understanding of the function and effect of genetic variants. This would result in obvious advantages in finding treatment of genetic diseases in humans but can also greatly impact animal/plant production and biotechnology. In this thesis, I have utilized known functional elements and other genome annotations to create models to predict the impact of genetic variants and utilized these models for the livestock species pig and chicken. These prediction models help to interpret the large amounts of data generated by modern massively parallel sequencing technologies as they steer the focus and research efforts to the most promising candidates, which otherwise would be indistinguishable from other candidates.

In animal breeding, quantitative geneticists are mostly interested in variants with predictive power for the phenotype of interest, independent of the exact role of the variant, as long as they co-occur with an actual functional variant. The consequent handling of the genome as a "black box" simplifies the methodology, because each allele is treated equally, but higher gains could be achieved in a more targeted approach based on a functional prioritisation of the variants. In the following sections, I discuss how our computational models can play a pivotal role in developing more targeted approaches for breeding and genome wide investigation of regulatory elements.

The focus of this discussion is on variant prioritisation tools that could in theory be applied for diploid species within the animal kingdom. CADD-like models can be generated for non-diploid / non-animal species, but their power have not been assessed in any study so far.

6.1. CADD-like models for non-human species can contribute to animal breeding

6.1.1. CADD-like models are suitable for variant effect prediction in animal genomes

When I tried to create a SNP prioritisation tool that can be used by researchers and breeders for genomic selection and the investigation of the genome, I was influenced by the CADD approach. This methodology, originally published in 2014 by Martin Kircher et al. [1], can generate a score for the putative deleteriousness of any SNP with respect to the reference genome. Deleteriousness is inherently linked to functionality and the expression of a phenotype, due to the changing ecological niche of a species which results in changing selection criteria over evolutionary time scales. Comparable scores for livestock would help in the breeding process and accelerate the elucidation of the coding and non-coding genome of livestock species, but these were not available before.

Before creating similar scores for livestock species I established, as an explorative step, the differences between humans and non-human species by creating a CADD-like model for mouse. I created several models, with decreasing numbers of genomic annotations, to study how models for livestock species, for which only a limited number of genomic annotations are available, may perform. CADD in its most recent version (v.1.5) uses 111 annotations for feature generation, many of which are not available for non-human species. Moreover, the lack of reasonably sized validation data sets and differences in the phylogeny of publicly available whole genome sequence alignments may have an effect on the capability of the scores to rank deleterious variations. To investigate these potential issues, I decided to conduct a feasibility study in mouse, as this model organism has relatively rich genomic annotations and SNP data sets with known functional effects

that can be exploited as validation sets. Our results have been presented in Chapter 2. The conclusion was that even with smaller genomic annotation resources, reasonable CADD-like models can be generated. These results motivated me to develop two CADD-like models for the livestock species pig (pCADD) and chicken (chCADD), which have been presented in Chapters 3 and 5. In Chapter 4, I have shown the capabilities of the CADD model for pig (pCADD) in identifying several novel, non-coding SNPs that would have been particularly difficult to identify without the use of pCADD. In the following sections, I will discuss several ways to use pCADD and chCADD scores to add value to existing approaches and issues.

6.1.2. Exploiting CADD to improve estimated breeding values

Estimated breeding values (EBV) are currently the basis for selection in animal breeding. Popular methodologies to calculate these are BLUP (Best linear unbiased predictor) approaches. CADD-like scores for livestock, such as chCADD and pCADD, could be directly applied to weight variants in BLUP approaches. A prominent BLUP algorithm is GBLUP [2] (Genomic Best linear unbiased predictor), which uses marker SNPs to compute a relationship matrix between phenotyped and non-phenotyped individuals. Based on these relationships, the individuals without a phenotype are assigned a weighted average phenotype score, derived from the phenotyped individuals. Commonly, the population allele frequency of the marker SNPs is included in the calculation of the relationship matrix to give a particular emphasis on low frequency alleles, with the underlying assumption that individuals which share alleles that are not common in the population should be more related. GBLUP in its original version [3] does not assume any additional weights, thus assumes the contribution of each SNP on a specified trait to be equal:

$$Z = M - P$$

$$G = \frac{ZDZ'}{2\sum p_i(1-p_i)}$$

where Z is the centered genotype matrix based on M, the n x m incidence matrix, with n genotyped individuals and m SNPs (coded as -1, 0 and 1) and P, a matrix of minor allele frequencies expressed as differences from 0.5. G is the genomic relationship matrix, with pi the minor allele frequency of SNP i. To add a weight, matrix D can be added to the equation as shown. D is a diagonal matrix containing the weights for SNP i. In the original GBLUP, the D matrix is an identity matrix, assigning each SNP the same weight. There are myriad of weighting strategies. These are often based on either predicted SNP effects, which are established by iteratively computing a G matrix and performing a whole genome regression to estimate SNP variances (WssGBLUP) [4], or by a summary of SNP effects, e.g. SNPs genomically in close proximity are assigned the same effect [5], [6]. pCADD or chCADD scores for each SNP can be used as weights in the D matrix. This would give an emphasis on variants that might be deleterious which are assumed to be more likely to have an impact on the phenotype.

Such an approach assumes that CADD-like scores are indicative for an effect on the expression of the investigated trait. This is not necessarily true, as CADD scores have not been developed with a particular trait in mind. CADD evaluates changes that may negatively affect the survival of the individual due to its assumptions on the deleteriousness of SNPs. Consequently, high scores (>25) are associated with lethal mutations, while scores in the range of ~15-25 may be associated with mutations that have a functional effect without being lethal/disease inflicting. Taking this into account, the pCADD and chCADD scores need to be adjusted to correlate better with the particular trait for which EBVs are calculated. Then the predicted phenotypes could be computed, with the assumption that individuals who share larger numbers of highly scored, trait specific alleles, will express more similar phenotypes.

One way of adjusting CADD-like scores for a particular trait is to exploit information about the LD structure of the alleles that are found significant in GWAS. It can be assumed that alleles with a high CADD-like score that are in high LD with the leading SNP from GWAS, are likely functional for the investigated trait and should result in more accurate EBVs. Another way to use CADD-like scores is to incorporate them directly in the GWAS calculations, to assign higher weight to SNPs that have a low allele frequency and would otherwise not show up in the analysis due to low statistical power.

6.1.3. CADD-like models can help to create population independent SNP arrays

SNP arrays provide a relatively cost-efficient way to genotype individuals, but they rely strongly on the LD structure of the investigated populations. LD is less prevalent in a population if it is genetically highly diverse. Moreover, LD is lost between populations if they are separated over longer evolutionary time spans. Thus, creating marker SNP arrays that are applicable for more than one target population becomes more difficult when linkage disequilibrium is not well preserved [7], [8]. Using causal variants on SNP arrays rather than SNPs expected to be in LD with the causal one would make the arrays more versatile and applicable for use on multiple populations as well as genetically highly diverse populations. In Chapter 5, I have shown that functionally causal variants can be identified with pCADD, which can serve as markers on a chip. CADD-like models can provide a similar framework for other species which would lead to more general SNP arrays, saving development cost and easing interspecies comparisons.

6.1.4. Potentially functional variants can act as proxies for phenotypes that are difficult to assess

CADD-like scores may also help in defining new proxies for phenotypes that are difficult to measure. For example, to determine resistance to bacterial/viral infections in livestock species, infection studies need to be conducted in many individuals to estimate the resistance of the population in different generations. In addition, the tested individuals need to be separated from the herd/stock, monitored, and, subsequently, culled due to the regulations on hygiene and animal experiments.

As an alternative approach, the presence of functional variants with a potential effect on the immune system and bacterial/viral resistance could serve as a proxy. Variant scores can be summarized to a phenotype score, meaning that individuals only need to be genotyped. Through GWAS on the target phenotype, we can retrieve, as described in Chapter 5, variants in high LD with the leading SNPs. Subsequently, we can investigate the effect of these SNPs on nearby genes and their gene products. After this is established, we can apply a Mendelian randomization approach in which we use the selected SNPs as instruments to compute the correlation of the gene products with bacterial/viral resistance. This will support the causality of these gene products on the phenotype. Moreover, the computed odds-ratio values can also be used as a proxy to infer the phenotype.

By pre-selecting SNPs, we can make assumptions about which genes may be affected by the likely causal SNPs and use them as intermediate phenotypes to compute a proxy for difficult to measure phenotypes.

6.1.5. Identifying mutations that can be introduced via genome editing

Genome editing has potential for the breeding industry as it can greatly increase the speed with which genomic variants can be introduced in a population [9]–[11]. However, while high precision genomic tools such as CRISPR/Cas9 [12] are available to make targeted edits, it is unclear what edits should be made to improve or realize specific traits. SNP prioritisation tools for livestock genomes such as pCADD and chCADD can help to identify potential functional candidates to introduce in the genome and their effect could be validated within one generation.

Even though such technology is achievable, applications in livestock breeding are currently neither supported by public nor by regulators [13], [14], even for applications that would notably improve animal welfare, a generally well supported cause [14], [15]. Currently, this prevents major developments and implementations of such breeding strategies [16] as minor bacterial DNA contaminations [17], through the molecular tools used, can lead to cancellation of large projects and total loss of investment. Until there is a shift in public opinion and regulations, funding for research and applications will likely be limited.

6.2. Usability of CADD-like scores for the evaluation of SNPs and regions

6.2.1. The evolutionary distance of the selected ancestor affects the number of derived variants

CADD-like scores have the benefit that they can, in theory, be generated for any species for which a sufficient number of close relatives are sequenced. Whole genome sequences are used to infer the genomes of ancestors of the species of interest at different evolutionary distances. In Chapter 5, we used the minimum number of sequences in the form of the 4-sauropsids EPO (Enredo, Pecan, Ortheus [18], [19] alignment from the Ensembl data base v.95. To generate the models in Chapter 2, 3 and 5, we employed the inferred ancestor sequence of the closest ancestor. This is not a necessity; even more distant inferred ancestral genome sequences can be used to derive variants. However, more distant ancestors will lead to less aligned genome sequence of the species of interest. On the other hand, a more distant ancestor sequence increases the number of derived alleles per aligned sequence due to larger differences between the sequences. Missing genome coverage is in principle not a problem for the generation of CADD-like models, as there is a surplus of training data. It could, however, alter the patterns learned during the training procedure when the loss of coverage is not uniform (see section 6.2.2 in this Discussion). Note that the learned CADD model can make predictions on unseen data, including regions of the genome for which no aligned sequences are available.

6.2.2. The underlying phylogeny can affect the grading of deleteriousness in CADD-like models

Depending on the depth of the underlying phylogeny, certain regions may be over- or underrepresented in the generated training set. An example is the difference in class distribution in the human and mouse training sets in Chapter 2 (Supplementary Table 2). Depending on the evolutionary distance between the species of interest and the other species considered, the distribution of SNPs in differently conserved regions in the training set could change. Genic regions may become overrepresented due to their generally higher conservation, leading to a larger

fraction of genic SNPs than would be expected based on the relative gene content in the genome of the species of interest.

Models generated from data derived from shallow phylogenies will be more precise in distinguishing between variants located in regions under recent selection, as compared to distinguishing between variants located in more variable regions. SNPs located in regions that have been under selection over longer time periods (thus lie in more conserved regions) would be generally considered deleterious, independent of their actual impact on function. On the other hand, if the phylogenies are extremely deep and cover large evolutionary distances between the different species, the emphasis may lay on the distinction between variants that are already located in very highly conserved regions. This would cause a weaker distinction among variants located in more variable or novel parts of the genome because they could be considered less impactful despite being functional.

The consequence of this effect is that the efficiency of evaluating variants depends on the underlying phylogeny and the evolutionary time frame considered. The evaluation of variants located in certain regions may be biased, either by being predicted to be more benign or more deleterious than they should be. That being said, this issue is to a certain extent being corrected by the computation of the log ranks of the model output rather than utilizing the posterior probability of the model directly.

6.2.3. Deeper phylogenies may lead to mis-estimation of substitution rates

For all CADD-like models that we have created, we used the variant simulator from the original CADD publication [1]. It uses the alignment of 4 genome sequences and their inferred ancestral sequences to derive substitution rates and simulate novel SNPs. The simulator creates a random number of novel SNPs by iterating over the genome and randomly deciding, based on the pre-derived mutation rates, if a certain position will be mutated. When a position is selected for mutation, the alternative allele is randomly selected based on the pre-derived substitution rates. The mutation and substitution rates are derived under parsimony [20] assumptions which neglect time and the possibility of multiple substitutions at the same site. This has several drawbacks, such as that ignoring the multiple substitution assumptions can affect phylogeny reconstruction [21], [22], resulting in mis-estimated substitution rates [23] important for the simulation of novel SNPs. Multiple substitutions can be ignored for shallow phylogenies but will become more severe for deeper phylogenies. A fixed threshold on phylogenetic distance to generate an optimal CADD-like model for the questions one wants to answer is hard to define and requires further study.

6.2.4. CADD-like scores should perform similar across different populations of the same species

Another open question is whether high-scoring variants always have a deleterious or otherwise functional effect in any subpopulation of the species of interest, or whether they differ between different subpopulations. This latter effect is also known for polygenic risk scores for humans that are often less accurate in human populations that are not of European ancestry [24], [25]. For the CADD-like models however, we do not expect major differences between subpopulations as these models learn patterns of putative deleteriousness from accumulated or presumably missing mutations over larger evolutionary time scales. As a result, the model will learn from variants that are deleterious throughout the entire species, and not from variants that arise from micro evolution within a species.

6.2.5. Comparability of CADD-like scores across different parts of the genome

In Chapter 3 I discussed the variant 3:43952776T>G in Sscrofa11.1. This variant has been shown to be lethal recessive [26], but scored relatively low with a pCADD value of 10.14. Hence, in a global genome-wide search for deleterious variants it may fall under the radar, because ~$7.1*10^8$ potential SNPs score higher than this lethal recessive allele. Compared to other potential SNPs in the same intron however, it is among the top 3, and it is the highest scored SNP for that intron occurring in the investigated populations.

Part of the reason for this behaviour is that CADD-like scores correlate positively with the number of genomic annotations available for a variant. Most often when genomic annotations are missing, these are missing for non-genic variants, leading to the situation that benign/non-functional variants in coding regions can have higher scores than functional or potentially deleterious variants in noncoding regions. This stands in conflict with the aim of creating a unified score for the entire genome, that can be applied to prioritize variants anywhere. Even between different genes, which are responsible for monogenic diseases, scores can vary significantly. This makes the application of a fixed threshold to define deleteriousness or functionality of variants difficult without considering any additional factors.

To address this issue, methods such as the mutation significance threshold [27] and Gavin [28] have been developed. These calculate variable, gene-specific thresholds based on variant data previously not used to create the CADD model. Unfortunately, the use of known functional variations as a baseline to evaluate the potential impact of novel mutations is not suitable for livestock species, due to a lack of such ground truth data. In addition, the focus on genes destroys the purpose of having a model capable of scoring variants in the entire genome.

Alternatively, a score could be developed that accompanies the CADD-score and is based on the number of non-imputed genomic annotations of a particular variant. It could be a sum of the feature weights, as they are applied by the model. In this way, missing a genomic annotation, coding for a model features with particular high weights, would cause a much larger change of the score than the imputation of several, less important genomic annotations. In this way the score could indicate how well two variants can be compared with each other. Incorporating the CADD-score and the score to measure comparability to one, would not be reasonable because then the meaning of the resulting score will be difficult to interpret.

6.3. Future of SNP prioritization

The field of functional effect prediction of variations in the genomes of human and non-human species will further develop, with the human genome as its trailblazer and spurred by the ever-growing number of data types and data sets. Models specific for one particular category of traits, populations or tissues can be created and may lay the foundation for ensemble approaches. The demand for scoring other types of variants such as structural variations (SV) is high but may still be far in the future due to the large number of different SVs and their various shapes and sizes.

Eventually, with a deeper understanding of the basic building blocks of gene function and regulation, effects of genotype on phenotype may perhaps be directly simulated, which would remove the need for independent SNP prioritisation.

6.3.1. Growing databases give the opportunity for more advanced genotype-phenotype predictions

From 2009 to 2015, the data storage capacity of the European Bioinformatics Institute (EBI) has increased by 1,150%, from 6 to 75 PetaBytes [29]. Besides whole genome DNA sequences for more and more species, the diversity of data is increasing as well. This is mirrored in the number of distinct and actively maintained molecular biological databases, of which there are 1,637 in January 2020 [30]. With the provision of data, it will become easier to develop new tools to prioritize variants for other species as well as develop new approaches. Richer data sets of functional variants will also help in the validation of these models, which so far has been a major obstacle for the development and validation of most models of non-human species.

Newly generated, massive data sets have already helped to develop a method addressing an issue that regular CADD-like models cannot address. Pleiotropic effects of SNPs, as reported in Chapter 4, are currently not considered in any way by CADD-like models due to their "one score represents all" approach. Xiang et al. [31] created so-called Functional-And-Evolutionary Trait Heritability (FAETH) scores for variants in the cattle genome. They derived the functional and trait heritability of variants for 34 phenotypes. To do so, the authors used functional annotations, GWAS and identified LD regions via conservation across several breeds in ~44,000 cattle. In total, they provided scores for 17.5 million variants, which is only a fraction of the bovine genome. Their general approach can be reproduced for any species and the methodology can be integrated in the creation of CADD-like models to emphasize pleiotropic SNPs.

6.3.2. Shallow phylogenies to create population and tissue specific CADD-like models

Biomarkers display different importance and predictability for an associated phenotype, depending on the subpopulations [32], [33]. This varying behaviour over subpopulations is not taken into account by CADD-like models. Even though, as discussed in section 6.2.4 of the Discussion, scores of CADD-like models are independent of individual population structure, it may be interesting for breeders to have a breed-specific CADD-like model which could help in identifying functional differences between their breeding lines. This could be done by using shallow phylogenies in which the ancestor sequences are not inferred between two different species but between two different subpopulations of the same species. Differences and similarities between models of the same species could help in pinpointing generally applicable biomarkers.

In theory, the idea of shallow phylogenies could be stretched even further: instead of exploiting the differences between subpopulations, it is imaginable to use differences between tissues to generate tissue-specific models, by exploiting accumulated somatic mutations and allele frequencies established through multiple single cell sequencing [34]. This could help to identify mutations detrimental for specific tissues. Results of already existing tissue specific predictors such as DeepBind, DeepSea, Basset [35]–[37] could be incorporated with an ensemble [38] algorithm to generate scores with a particular emphasis on tissue specificity.

Even though both applications of shallow phylogenies may be valid, at least the tissue specific model might be problematic to construct. In both cases, the evolutionary time scale may be too short to properly deplete the reference genome of the subpopulation from deleterious variants to display meaningful patterns from which a training algorithm could learn. With careful selection, a subpopulation-specific CADD-like model may still be possible; tissue-specific models are likely impossible because they rely on a sufficient number of somatic mutations. Considering that the average mutation rate per cell division in humans [39] is between 2.4 to 29.6 x 10^{-7}, identifying enough somatic mutations will be difficult, even with modern, massively parallel single cell

sequencing approaches that sequence hundreds of thousands of cells simultaneously. In summary, it may be useful to incorporate predictions from tissue-specific models to the set of features for a CADD-like model but creating tissue-specific CADD-like models is not trivial.

6.3.3. Tools for structural variation prioritisation

Among the prioritisation tools created so far, tools to prioritise structural variations (SVs) are underrepresented. Until recently, with the occurrence of long read sequencing technologies, it has been difficult to accurately detect SVs and/or their exact locations. To prioritise SVs it is paramount to evaluate or predict the effect they have on the phenotype of individuals carrying the mutation. There is no doubt that they can influence function [40], [41]. To date, a large number of algorithms have been developed that are able to detect various classes of SVs [42] but the impact on function and the effect on the phenotype remains difficult to estimate [41], [43], [44] and mostly involves association with expression quantitative trait loci (eQTL) or overlap with other known functional elements [44], [45]. To the best of my knowledge, there is only one method available that quantifies the predicted impact of SVs, SVScore [46]. SVScore utilizes the per-base, genome-wide available CADD scores to aggregate them for intervals specifically defined for each structural variant, but it does not address gene fusions or novel adjacencies with cis-regulatory elements. A similar method could be easily created for the investigation of SVs in non-human species, given that a CADD-like score is available.

SVs do have the inherent difficulty that they can occur in various sizes and classes. Moreover, as for SNPs, annotated variants hardly exist for non-human species. ClinGen [47] is a manually curated data base with 59,71349 [48] clinically annotated SVs (accessed 04-02-2020) for human. Databases such as these form the basis for our understanding of SVs and should be set up for animals as well if we want to make progress in this area.

6.4. Final remarks

With this thesis, I have laid the foundation for in-silico DNA sequence variant effect predictions in the genomes of livestock species. Utilizing a well-established method for variant prioritisation in human genomes, I have shown that for livestock species, similar methods can be created. By deploying such models, I believe functional variants will be easier identified and that there is a bright future for the field of animal genetics and breeding.

Bibliography

[1] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants.," *Nat. Genet.*, vol. 46, no. 3, pp. 310–5, 2014.

[2] T. Meuwissen, B. Hayes, and M. Goddard, "Genomic selection: A paradigm shift in animal breeding," *Anim. Front.*, vol. 6, no. 1, pp. 6–14, 2016.

[3] P. VanRaden *et al.*, "Invited review : Reliability of genomic predictions for North American Holstein bulls," *J. Dairy Sci.*, vol. 92, no. 1, pp. 16–24, 2009.

[4] H. Wang, I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir, "Genome-wide association mapping including phenotypes from relatives without genotypes," *Genet. Res. (Camb).*, vol. 94, no. 2, pp. 73–83, 2012.

[5] X. Zhang, D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal, "Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS," *Front. Genet.*, vol. 7, no. AUG, pp. 1–14, 2016.

[6] M. Teissier, H. Larroque, and C. Robert-Granié, "Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative trait influenced by a major gene," *Genet. Sel. Evol.*, vol. 50, no. 1, pp. 1–12, 2018.

[7] Y. B. Fu, M. H. Yang, F. Zeng, and B. Biligetu, "Searching for an accurate marker-based prediction of an individual quantitative trait in molecular plant breeding," *Front. Plant Sci.*, vol. 8, pp. 1–12, 2017.

[8] H. Sun *et al.*, "Genome-wide and trait-specific markers: A perspective in designing conservation programs," *Front. Genet.*, vol. 9, pp. 1–4, 2018.

[9] R. L. Gratacap, A. Wargelius, R. B. Edvardsen, and R. D. Houston, "Potential of genome editing to improve aquaculture breeding and production," *Trends Genet.*, vol. 35, no. 9, pp. 672–684, 2019.

[10] S. Gonen, J. Jenko, G. Gorjanc, A. J. Mileham, C. B. A. Whitelaw, and J. M. Hickey, "Potential of gene drives with genome editing to increase genetic gain in livestock breeding programs," *Genet. Sel. Evol.*, vol. 49, no. 3, pp. 1–14, 2017.

[11] Y. Zhang, K. Massel, I. D. Godwin, and C. Gao, "Applications and potential of genome editing in crop improvement," *Genome Biol.*, vol. 19, no. 210, pp. 1–11, 2018.

[12] F. Zhang, Y. Wen, and X. Guo, "CRISPR / Cas9 for genome editing : progress , implications and challenges," *Hum. Mol. Genet.*, vol. 23, no. 1, pp. 40–46, 2014.

[13] E. Callaway, "EU law deals blow to CRISPR crops," *Nature*, vol. 560, no. 16, 2018.

[14] M. C. Yunes, D. L. Teixeira, M. A. G. von Keyserlingk, and M. J. Hötzel, "Is gene editing an acceptable alternative to castration in pigs ?," *PLoS One*, vol. 14, no. 6, pp. 1–18, 2019.

[15] D. F. Carlson *et al.*, "Production of hornless dairy cattle from genome-edited cell lines," *Nat. Biotechnol.*, vol. 34, no. 5, pp. 479–481, 2016.

[16] H. Ledford, "Creators of gene-edited animals bypass US market," *Nat. Biotechnol.*, vol. 566, pp. 433–434, 2017.

[17] A. L. Norris, S. S. Lee, K. J. Greenlees, D. A. Tadesse, M. F. Miller, and H. A. Lombardi, "Template plasmid integration in germline genome-edited cattle," *Nat. Biotechnol.*, vol. 38, pp. 163–164, 2020.

[18] B. Paten, J. Herrero, K. Beal, S. Fitzgerald, and E. Birney, "Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs," *Genome Res.*, vol. 18, no. 11, pp. 1814–1828, 2008.

[19] B. Paten *et al.*, "Genome-wide nucleotide-level mammalian ancestor reconstruction," *Genome Res.*, vol. 18, no. 11, pp. 1829–1843, 2008.

[20] E. Sober, "Explanation in biology : Let's razor Ockham's Razor," *R. Inst. Philos. Suppl.*, 1990.

[21] K. M. Halanych and T. J. Robinson, "Multiple substitutions affect the phylogenetic utility of Cytochrome b and 12S rDNA Data: Examining a rapid radiation in Leporid (Lagomorpha) evolution," *J. Mol. Evol.*, no. 48, pp. 369–379, 1999.

[22] J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach," *J. Mol. Evol.*, no. 17, pp. 368–376, 1981.

[23] R. S. Schwartz and R. L. Mueller, "Variation in DNA substitution rates among lineages erroneously inferred from simulated clock-like data," *PLoS One*, vol. 5, no. 3, 2010.

[24] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, "Current clinical use of polygenic scores will risk exacerbating health disparities," *Nat. Genet.*, vol. 51, no. 4, pp. 584–591, 2019.

[25] A. R. Martin *et al.*, "Human demographic history impacts genetic risk prediction across diverse populations," *Am. J. Hum. Genet.*, vol. 100, no. 4, pp. 635–649, 2017.

[26] M. F. L. Derks *et al.*, "Loss of function mutations in essential genes cause embryonic lethality in pigs," *PLoS Genet.*, vol. 15, no. 3, pp. 1–22, 2019.

[27] Y. Itan *et al.*, "The mutation significance cutoff: gene-level thresholds for variant predictions," *Nat. Methods*, vol. 13, no. 2, pp. 109–110, 2016.

[28] K. J. van der Velde *et al.*, "GAVIN: Gene-Aware Variant INterpretation for medical sequencing," *Genome Biol.*, vol. 18, no. 1, pp. 1–10, 2017.

[29] C. E. Cook, M. T. Bergman, R. D. Finn, G. Cochrane, E. Birney, and R. Apweiler, "The European Bioinformatics Institute in 2016: Data growth and integration," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D20–D26, 2016.

[30] D. J. Rigden and X. M. Fernández, "The 27th annual Nucleic Acids Research database issue and molecular biology database collection," *Nucleic Acids Res.*, vol. 48, pp. 1–8, 2020.

[31] R. Xiang, I. Van Den Berg, I. M. Macleod, B. J. Hayes, and C. P. Prowse-wilkins, "Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits," *Proc. Natl. Acad.*

Sci. U. S. A., vol. 116, no. 39, 2019.
[32] A. Jemal, J. Lortet-tieulent, E. Ward, J. Ferlay, O. Brawley, and F. Bray, "International variation in prostate cancer incidence and mortality rates," *Eur. Urol.*, vol. 61, pp. 1079–1092, 2012.
[33] S. Enroth, A. Johansson, S. B. Enroth, and U. Gyllensten, "Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs," *Nat. Commun.*, vol. 5, no. 4684, pp. 1–11, 2014.
[34] L. Zhang, X. Dong, M. Lee, A. Y. Maslov, T. Wang, and J. Vijg, "Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 18, pp. 9014–9019, 2019.
[35] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.
[36] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model.," *Nat. Methods*, vol. 12, no. 10, pp. 931–4, 2015.
[37] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Res.*, vol. 26, no. 7, pp. 990–999, 2016.
[38] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, 2010.
[39] D. J. Araten *et al.*, "A quantitative measurement of the human somatic mutation rate," *Cancer Res.*, vol. 65, no. 18, pp. 8111–8117, 2005.
[40] J. R. Lupski, "Genomic rearrangements and sporadic disease," *Nat. Genet.*, vol. 39, no. 7S, pp. S43–S46, 2007.
[41] D. M. Bickhart and G. E. Liu, "The challenges and importance of structural variation detection in livestock," *Front. Genet.*, vol. 5, pp. 1–14, 2014.
[42] S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani, "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing," *Genome Biol.*, vol. 20, no. 1, pp. 8–11, 2019.
[43] L. Han *et al.*, "Functional annotation of rare structural variation in the human brain," *Eur. Neuropsychopharmacol.*, vol. 29, 2019.
[44] P. H. Sudmant *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015.
[45] H. Yang and K. Wang, "Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR," *Nat. Protoc.*, vol. 10, no. 10, pp. 1556–1566, 2015.
[46] L. Ganel, H. J. Abel, and I. M. Hall, "SVScore: An impact prediction tool for structural variation," *Bioinformatics*, vol. 33, no. 7, pp. 1083–1085, 2017.
[47] H. L. Rehm *et al.*, "ClinGen — The Clinical Genome Resource," *N. Engl. J. Med.*, pp. 1–8, 2015.
[48] "dbVAR - Structural Variation Data Hub." [Online]. Available: https://www.ncbi.nlm.nih.gov/dbvar/content/human_hub/#clinical. [Accessed: 04-Feb-2020].

Acknowledgements

With the acknowledgements my thesis ends and over 4 years of PhD education. Even though on the clock only 4 years have passed, looking into the mirror, my grey hair and growing winkles show this Ph.D. did not just cost me these 4 years. Perhaps, writing up this section adds even more grey hair because of the difficulty to give sufficient credit to everyone who accompanied me along this journey.

First, I would like to give my gratitude to my promotors **Marcel** and **Dick**. Without them this thesis would not have come together, and I am thankful for all the effort they put into me even though I may have cost them also more than just 4 years. **Marcel** I am always surprised by how quick responds I can get from you, no matter how late it is. Sometimes I was a bit confused about what your real opinion on a matter was until I realized that you have a passion for debates and for the sake of discussion you would defend a position which may not be yours. Truly a valuable skill to have for an academic. **Dick,** thank you for giving me so many constructive suggestions for my Ph.D. research they always brought me back on track. You deserve all my appreciations and respect as how much effort you invested in me.

Second my collaborators from the Breed4Food consortium, **Martijn**, **Hendrik-Jan**, **Mirte** and **Martien**, it was always pleasant to hear about the challenges and prospects in the animal breeding world. Unfortunately, waaay to late I actually went to a proper animal genomics conference (ISAG, 2019) which helped me greatly to put my own research in perspective. My eternal feasibility study did not help with that either. But as soon as that was sorted out the real work could be kicked off and for that I am extremely thankful to you **Martijn**. After waiting for more than 2 years and only hearing about one problem after the other during our regular lunch meetings, you immediately saw the potential of our combined expertise and we set a study in motion which I think has been pretty successful. **Chiara** not strictly a collaborator from the Breed4Food consortium, nevertheless another hard-working PhD who is passionate about animal genetics and motivated the investigation of conserved regions via chCADD. I hope all the best for your own defense and that the sun is not turning you into a crab like it did to me.

As a sandwich PhD, between Delft and Wageningen, I had always the blessing to be able to join two group outings a year, rather than only one. Therefore, my connection to both groups feels rather strong even though I neglected more and more my Wageningen group to the end of my PhD, especially since my accident. To show my gratitude I want to give special thanks to all the native Spanish speakers of that group **Miguel**, **Victoria** and **Hernando**. When we travelled back together to Utrecht, it was always a blast. I will remember **Miguel** as the only other person who I know who actually played the trilogy of the Shadow Run video games, @**Hernando**, Marcela is currently next to me and still growing strong, she even blossomed recently. Unfortunately, the non-native Spanish speakers of the group are still in majority and cannot all be named individually. Nevertheless, I want to try at least to give some kind of a sample. Depending on where I was sitting, my neighbors were either **Raul** or **Vittorio**. Two completely different characters, neither I wanna miss. **Raul** will take over the Bioinformatics group at some point when Dick is going to retire, and **Vittorio** will have left Bioinformatics at that point because he made a fortune as a sausage seller in Germany.

Coming to my other group of the last 4 years, **Pattern Recognition & Bioinformatics**, a group consisting in itself of 4 groups which (at least the Bioinformatics part) can be partitioned further, depending on the day time and the mood of its members. Great Research icons such as the heads of the Plant Deep Learning group or Computational Microbiology were among those which established their own realm there. Independent of the actual background, all theoretically united in applying pattern recognition techniques to answer biological questions. Over the years, we moved

from the high-rise to the new building (whatever the name of it is) and even before the move, the composition of my roommates constantly changed until I had a comfortable private office at some point. Then, a couple of weeks before we moved, **Soufiane** joined the group and was placed in my office. After the move in the new building, everything seemed fancier but less practical. Especially toilets, plentiful before, were now few and often broken. I cannot help myself but with **Soufiane** there is the constant subtle association that with the coming of the French, German prosperity declined. Another new roommate, **Tom**, was actually an old roommate, before he left me for a more beautiful one with almost the same name. **Christine**, due to your contra-anal attitude you are always successful in brighten up the work atmosphere and a reliable source of distraction in case that the work routine has become too mundane. The next in line is **Stavros**, I am not sure anymore if this has always been the case but definitely after moving to the new building, he became a greek sports god with a particular interest in basketball. An easy-going person to have jokes with no matter if beers are involved or not, if the jokes are good or not or even politically correct. The majority of my time in the PRB group I definitely spend with a real **OOG**. A person whose deeds are too long to count and who was born with a name different to the one he is currently carrying. A person who even makes my wife jealous because she is constantly surrounded by his name in different forms, such as Half-Life **Alyx** or my sports bag **Alex** or because of the many borrel stories I have told her about him or because he was actually there, as a best man, when I married her. A person I don't wanna miss and besides all of that, actually has some proper knowledge about Bioinformatics and big shots in the field. The last two people in the Partyroom 5.920 are **Mustafa** and **Aysun**. As the newest additions to the Bioinformatics group, my shared history with them is of course the shortest which is kinda sad. I guess there could have been great potential to write epic stories in future because both of them harbor much potential on a personal and professional level.

Now I am already writing Acknowledgements for 1 ½ pages and still I haven't even finished the entire Delft Bioinformatics group yet. One of my earliest encounters in the Delft Bioinformatics Group was with **Joanna** who as a PostDoc was located in my office and then became Assistant Professor due to her persistent hard work. Together with **Arlin**, I applied to Joanna's PostDoc position. Eventually, for none of us did this work out but with **Arlin** I have many more memories than just competing for the same position. Memories about an extremely fun person who also has a weak spot for metal music. **Tamim**, similar to **Stavros**, always tried to motivate me to play with them either football or basketball, considering the number of injuries in our group I am quite glad that I successfully resisted until the end :P. **Meng** and **Ramin**, two more of Thomas' PhD students. I guess **Ramin** is missing the office the most, before Corona was a thing, **Ramin** had great ambitions and wanted to take the stairs every single time, now I guess he has no access to too many stairs anymore. **Meng** always thoughtful and considerate in all your actions. As someone from 南京市 I am always curious how you perceive being in a country with only twice as many people?

Tom, I am not sure where to start, the first time we've met we met when you were explaining to me your master thesis about active learning and how I may be able to make use of it in my own research. Since then your voice has been a stable force at every borrel and everywhere, where people were able to hear it, there was immediately joy in the air and noise complaints by less fun people. **Marco** it is a mystery for me how your body works but I am glad that I was able to meet you and that you were among the few in Copenhagen. A scientist through and through without the any obnoxious attitude of superiority. Especially recently I learn to value your attitude and propositions more and more 'Many scientists take themselves too serious but their research not serious enough.' **Ekin** and **Laura** as siblings in mind, you have to share these few words. I think over the course of my time here, I have not been at anyone's place more often than at Ekin's but I also have never been so late at anyone's than at Ekin's. Also, I think I never said sorry for the Kapsalon, so I do it here and now 'I am sorry'. **Laura**, with all rights proud to be an engineer and

always around to take record of other people's misdemeanours. After you moved back to your "*island"* I really felt that things are changing now. I hope that we will be able to visit you sometime in future. Another person who already left us earlier, **Wouter**, must not be forgotten, after all I am still living in the apartment he left behind, even though it seems that my time here is approaching its end. **Wouter** was among the first of the PR group I actually met, of course during a borrel, but unlike **Alex**, my first impression of him was rather civilized. It still amazes me how time and being a daddy can tame even the most unbound force. But eventually we will all experience that I think. A lot of thanks for the laughs and many suggestions. I will never forget the tractor channel (very important).

Und nun nach über 2 Seiten kommen wir endlich zu dem Deutschen Teil der Acknowledgements. **Alexander**, als Marco uns gefragt hat was wir über dich sagen könnten fiel es mir echt schwer irgendetwas Peinliches zu finden, ein Musterbeispiel eines Doktoranden, der auch ganz ohne Alkohol an borrels Spaß haben konnte. Ich wünsch dir noch alles Gute unabhängig davon was noch auf dich zukommt und natürlich viel Spaß bei Cyberpunk 2077.

Oft kommen die **Eltern** und andere **Familienmitglieder** an erster Stelle in den Acknowledgements, sorry, dass Ihr so lange blättern musstet, um hierhin zu gelangen. Ich wollte euch natürlich auch danken, da Ihr mir die Möglichkeit geboten habt, dass ich überhaupt erst im Ausland studieren konnte, sei es in den Niederlanden oder in Großbritannien. Ohne Unterstützung von zu Hause wäre das wohl so nie möglich gewesen. **Stephan** obwohl, oder wahrscheinlich gerade, weil wir unterschiedliche Bildungsrichtungen eingeschlagen haben nach unserem Abitur, hast du mir auch andere Sichtweisen nahegelegt, jedes Mal, wenn wir uns bei den Eltern gesehen haben und wir höchst Politische Themen angesprochen haben. **Corinna, Konni, Josephine & unbekanntes Geschwisterchen*** unabhängig davon welches Thema oder einfach nur mal quatschen, bei euch hat man immer ein offenes Ohr oder wenn man mal kurzzeitig beim VR Spielen einspringen muss. Ich wollte auch Josi für die vielen herzzerreißenden Videos von einem lachenden Baby danken und auch wenn noch nicht vorhanden, für zukünftige Videos dem unbekannten Geschwisterchen. Omas **Grete** & **Bärbel** und Opa **Jochen**. Mein Dank gilt auch euch für die vielen Kindheitserinnerungen und Erklärungen über die Welt. Wir sind auch nur die Summe unserer Erfahrungen und Genetik und ihr hattet zu beidem einen erheblichen Anteil, der dann in diesem Manuskript zusammengekommen ist.

Deutsch aber nichtmehr offiziell Familie, **Arthur**, **Markus** und **Thomas**, als meine ältesten Freunde hatten wir über all die Jahre in denen ich schon nichtmehr in Troisdorf/Bonn gewohnt habe immer noch Kontakt miteinander und das bedeutet mir doch recht viel da man auch Kontakte außerhalb seiner Arbeitsstelle braucht und man im Alter immer weniger Kontakte hat und wenn neue hinzukommen, die sehr häufig nur aus der Arbeitswelt und sehr oberflächlich sind. **Patrick**, als mein Bachelor Supervisor und später in England Projektleiter. Jemand der auch prima in die Gruppe von anderen Chaoten gepasst hätte, die oben bereits erwähnt wurden. Du hast auch eine innige Liebe für die Wissenschaft ohne Eitelkeiten was ich massiv bewundere. Leider sind unsere Themengebiete auseinandergedriftet, aber wer weiß wie man nochmal zusammenkommen kann.

段风梅，路增祥是我来自另一个家庭的爸爸妈妈。尽管我们物理距离很遥远，你们依旧欢迎我成为家人并支持我们。为此，我们永远感激不已。 很抱歉，现在您不能与我们在一起，但我们尽快会再次见面。

大宝宝（路璐），我们在火车上相遇。 我从来没奢想过你会成为我的妻子，也从来没想过我们在一起会笑着么多。不久，我们会住在自己的家中。 正如你常说的，我们的未来会越来越好。 无论我们在荷兰，德国，欧洲还是中国，我们都将为小宝带来更美好的未来。 和你以往的日子很美好，我们未来的日子会更美好的。为此，我会永远感谢你，当然也感谢你做的饭 :P

Curriculum Vitæ

Christian Gross was born 11th November 1988 in Hoexter, Germany. 2009 he received his ISCED 3 degree at Gymnasium Zum Altenforst in Troisdorf. His graduation was soon followed by a 14-month military service in the German Armed Forces at the Streitkräfteunterstützungskommando ABC-Abwehr und Schutzaufgaben in Cologne – Wahn, during which he received special distinction for the identification of an error in the ammunition storage guidelines. Following his military service, he pursued a Bachelor of Science in Biology at the University of Bonn which he graduated from 2013. During his Bachelor education he worked as a math teacher for Biology freshman and followed up on extracurricular courses/activities in Bionics and C++ programming. His Bachelor Thesis "*Testing new phylogenetic algorithms using LoBraTe – new developed process pipeline for comprehensive comparative tree reconstruction analyses*" was written at the Natural History Museum Bonn, Germany. Christian decided to follow up his B.Sc. Biology degree with a joint master program in Bioinformatics by the University of Amsterdam and Free University of Amsterdam (VU Amsterdam). In 2015 he received his M.Sc. (cum laude) from the VU. While he was conducting his M.Sc. studies he participated in the "Roche & OBR: Game Changing Innovation" competition and did internships at the Medical Research Council – Biostatistics Unit in Cambridge, United Kingdom and Natural History Museum in London, United Kingdom. In Cambridge he worked on drug response prediction, in London on phylogenetic tree reconstructions and validations. Beginning 2016 he joined the Pattern Recognition & Bioinformatics Group at the TU Delft and the Bioinformatics Group at the Wageningen University & Research to conduct his PhD study in predicting variant effects in the genomes of livestock species. Since May 2020 he is working as a PostDoctoral Researcher at the University Medical Center Utrecht, the Netherlands and University College London, United Kingdom in the field of genetically guided drug development.

List of Publications

First author

- Martijn F. L. Derks, Christian Groß, Marcos S. Lopes, Marcel Reinders, Mirte Bosse, Arne B. Gjuvsland, Dick de Ridder, Hendrij-Jan Megens, Martien A.M. Groenen, "Accelerated discovery of functional genomic variations in pigs" (submitted)

- Christian Groß, Chiara Bortoluzzi, Dick de Ridder, Hendrik-Jan Megens, Martien AM Groenen, Marcel Reinders, Mirte Bosse, "Evolutionarily conserved non-protein-coding regions in the chicken genome harbor functionally important variation", bioRxiv 2020.03.27.012005; doi: https://doi.org/10.1101/2020.03.27.012005

- Christian Gross, Martijn FL Derks, Hendrik-Jan Megens, Mirte Bosse, Martien AM Groenen, Marcel Reinders and Dick de Ridder, "pCADD: SNV priortisation in Sus scrofa", Genetics Selection Evolution 2019

- Christian Groß, Dick de Ridder and Marcel Reinders, "Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse", BMC Bioinformatics 2018, 19:3732.

Co-author

- Martijn FL Derks, Hendrik-Jan Megens, Mirte Bosse, Jero Visscher, Katrijn Peeters, Marco CAM Bink, Addie Vereijken, Christian Gross, Dick de Ridder, Marcel JT Reinders, Martien AM Groenen, "A survey of functional genomic variation in domesticated chickens", Genetics Selection Evolution 2018, 50:17 2.

- Patrick Kück, Mark Wilkinson, Christian Groß, Peter G Foster, Johann W Wägele, "Can quartet analyses combining maximum likelihood estimation and Hennigian logic overcome long branch attraction in phylogenomic sequence data?", PLoS ONE 2017, 12(10)

- Patrick Kück, Sandra A Meid, Christian Groß, Johann W Wägele, Bernhard Misof, "AliGROOVE–visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support", BMC Bioinformatics 2014, 15:29