Investigation of GWAS variants associated with loin depth in commercial pigs

S. Desire[1*], M. Johnsson[1,2], R. Ros-Freixedes[3], C. Chen[4], W. Herring[4], G. Gorjanc[1], R. J. Mellanby[5], J. Hickey[1], M. Jungnickel[1]

[1]The Roslin Institute, The University of Edinburgh, Midlothian, UK; [2]Swedish University of Agricultural Sciences, Uppsala, Sweden; [3]Departament de Ciència Animal, Universitat de Lleida, Agrotecnio (CERCA Center), Lleida, Spain; [4]The Pig Improvement Company, Genus plc, Hendersonville, TN, USA; [5]The Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, UK; *sdesire@ed.ac.uk

Abstract
Genome wide associate studies (GWAS) aim to identify genomic regions implicated in phenotype expression, but identifying true causative variants is difficult. Pig combined annotation dependent depletion (pCADD) scores may provide a straight-forward method of prioritising causal variants identified by GWAS. This study aimed to test using pCADD to identify candidate variants using loin depth in four pig lines. Three main regions on SSC7, 9, and 16 were identified in both maternal and terminal lines and have been found to be associated with carcass traits in previous studies. Two genomic regions were identified in maternal lines only, with SNPs within *MC4R* and *PIK3C2A* identified as the most likely causative variants. Several of the candidate SNPs have plausible biological links to muscle-related phenotypes, however further work is required to determine whether these are true causal variants.

Introduction
In genome-wide association studies (GWAS), genetic variants are independently tested for statistical association with a phenotype, with the aim of identifying the underlying causal genetic basis of the trait. Identifying true causal variants is difficult, due to high linkage disequilibrium (LD) between variants, a dilemma that is exacerbated in livestock species that have undergone selection for specific traits. While fine-mapping aims to solve this problem, there are many methods to choose from, with varying degrees of complexity. Pig combined annotation dependent depletion (pCADD) is an algorithm that estimates the deleteriousness of a given variant in the pig genome. As demonstrated by Derks *et al.* (2021), incorporating pCADD scores in the GWAS pipeline may provide a simple method for identification of candidate causative SNPs. The objective of this study was to use GWAS and pCADD to computationally identify SNPs that are likely to contribute to variance in pig loin depth, and to explore their biological plausibility by examining the genes in which these SNPs are located.

Materials & Methods
Animals and Phenotypes. Analyses were performed on 232,057 pigs from the Pig Improvement Company (PIC; Hendersonville, TN). Pigs from two maternal lines (recoded as A and B) and two terminal lines (recoded as C and D) were used in this analysis. Breeding values for loin depth measured via ultrasound were calculated as part of routine genetic evaluations. Breeding values were de-regressed (dEBV) as per VanRaden & Wiggans (1991), and used as phenotypes in GWAS. To account for variation in the reliability of dEBVs, a weighting value based on Garrick *et al.* (2009) was calculated for each individual, and the reciprocal of each weight was used as a residual weighting factor in GWAS.

Genotypes and quality control. Tissue samples were acquired from tails, which were docked within 48 hours of birth, and samples were genotyped using a customized Illumina single-nucleotide polymorphism (SNP) chip, mapped to susScr11.1. SNPs with MAF < 0.01 or a call rate < 0.95 were removed from further analyses. After quality control, an average of 42,370 (SD 1,514) SNPs were available for each line and used in GWAS analysis. Variants from whole genome sequence datasets were available, imputed via hybrid peeling (Ros-Freixedes *et al.,* 2020), and these SNPs were used to identify candidate SNPs, post-GWAS.

Association analyses. GEMMA software (Zhou & Stephens, 2012) was used to estimate centered genomic relatedness matrices. GEMMAs was also used to apply a series of univariate linear mixed models:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{e}, \quad (1)$$

where $\mathbf{y}$ is the vector of de-regressed EBVs, the vector $\boldsymbol{\beta}$, represents the vector of effect size of each SNP, and $\mathbf{X}$ is the incidence matrix relating genotypes with $\boldsymbol{\beta}$. The vector $\mathbf{u} \sim \mathbf{N(0, G_c\sigma_u^2)}$ represents polygenic additive effect with covariance equal to the centred genomic relationship matrix $\mathbf{G_c}$ and variance $\boldsymbol{\sigma_u^2}$. The vector $\boldsymbol{e}$ represents residual error, weighted to account for the variable reliability of de-regressed breeding values. To conservatively account for multiple comparisons in the GWAS, P-values were adjusted by Bonferroni correction. A SNP was significant at the genome level if -log10(p value) was greater than -log10(0.05/n), where n represents the number of markers across the genome.

Identifying candidate SNPs. A region spanning 0.5 Mbp up and downstream of each significant SNP was defined as a distinct genomic region. Significant SNPs within 0.5 Mbp of each other were merged into a single region. The GWAS SNP with the highest significance from each genomic region (lead SNP) was identified, and all SNPs in high LD (≥0.7) with lead GWAS SNPs were identified from sequence data. We obtained pCADD scores (Groß *et al.,* 2020) for each of these SNPs, along with their predicted effect, using Ensembl Variant Predictor (VEP). The variant with the highest pCADD score found to be in high LD with a lead SNP is referred to as the seqSNP. The genes harbouring seqSNPs were identified from the Ensembl database.

Results and discussion

The aim of this study was to identify regions associated with loin depth in four pig lines, and to interrogate the candidate causative SNPs, as suggested by pCADD scores. Table 1 lists the strongest associations found within each genetic line, along with a corresponding seqSNP for each region. While SNPs with strong genome-wide associations with loin depth were found for all genetic lines, no common regions were found across all four lines (Figure 1). Of the exonic seqSNPs identified, 11 were missense variants, 8 were synonymous, and one stop-gain. Six of the seqSNPs were intronic and 5 were intergenic.

One genomic region on SSC7 (~30Mbp) was found to be statistically associated with loin depth in both maternal lines, and one terminal line (D). This region has been previously associated with loin muscle area and other body traits in pigs (Zhuang *et al*., 2019; Bian *et al.*, 2021). Three separate seqSNPs were identified by pCADD scores for this region. Two were missense SNPs located in *GRM4* and *FANCE,* and one was a synonymous variant located in *CPNES*.

One maternal line and both terminal lines shared a large genomic region located on SSC16 (~32 – 34 Mbp) in common. Two of the seqSNPs identified for this region were located within known genes *ITGA1* and *ARL15*, while the third was an intron variant located in *MAST4*. Several studies have found this SSC16 region to be associated with loin muscle (Zhuang *et al.*,

2019; Bergamaschi *et al.*, 2020) and other body and growth-related traits (Gozalo-Marcilla et al., 2021). In addition, *ITGA* has been shown to be up-regulated in steers with low marbling in the *Longissimus Dorsi* muscle (de las Heras-Saldana *et al.*, 2020). The gene *GRIK4* located on SSC9 (~47.7 Mbp) was implicated in one maternal and one terminal line and has been previously associated with loin muscle depth in pigs (Bian *et al.,* 2021). *MAST4* has not previously been implicated in pig production.

The maternal lines shared two significant genomic regions on SSC1 (~160.7 Mbp) and SSC2 (~41.5 Mbp). The most likely seqSNP identified for these regions within both breeds were missense variants located within the *MC4R* (SSC1), and *PIK3C2A* (SSC2). *MC4R* is a known regulator of feed intake, growth and meat quality traits (Kim *et al.*, 2000). Derks *et al.*, 2021 predicted a missense mutation within *MC4R* to be associated with fat deposition and growth in pigs, while (Reyer *et al.,* 2017) found a SNP near *MC4R* to be statistically linked to lean mass percentage and growth. On SSC2, *PIK3C2A* has been linked to lean meat percentage in pigs (Zhou *et al.*, 2021). Cánovas *et al.* (2010) showed *PIK3C2A* to be downregulated in pigs in a high lipid deposition group, while *PIK3C2A* is up-regulated in the high-marbled Hereford cattle (Roudbari *et al*., 2020).

The strongest signals in the present study have been replicated elsewhere, supporting the evidence that these regions are important to the genomic control of muscle traits. pCADD identified several candidate SNPs, some of which were in genes that have plausible biological links to muscle phenotypes, although some were located a considerable distance away from the lead SNP. It is notable that a large proportion (35%) of variants flagged by pCADD in this analysis were missense mutations. CADD was developed to estimate the pathogenicity of SNPs in humans (Kircher *et al.,* 2014), and as such prioritises potentially deleterious variants with protein-coding consequences. Generally, GWAS variants are expected to have small effect sizes, therefore the genes they impact are likely to have a small effect on the phenotype. Further work is required to determine whether the missense variants prioritised in this study have a functional effect on muscle phenotypes. Several intergenic, intronic, or synonymous variants were predicted to be likely candidate variants by pCADD scores. The next step is to perform eQTL analysis on porcine muscle sampled from the genetic lines used in the present study. These data will aid the interrogation of candidate causal SNPs.

Table 1. The most highly significant GWAS associations (lead SNPs) for de-regressed EBVs for loin depth in each of four purebred pig lines, and candidate causative SNPs (seqSNPs), identified using pCADD scores.

		lead SNP		seqSNP		
line	SSC	Position (Mbp)	GWAS - log10pvalue	Position (MbP)	pCADD score	Predicted consequence
A	6	45.74	11.14	52.0	34.95	missense
B	2	0.40	18.83	0.8	41.97	synonymous
C	16	34.10	21.55	34.0	43.19	missense
D	5	66.10	27.4	66.2	15.82	intergenic

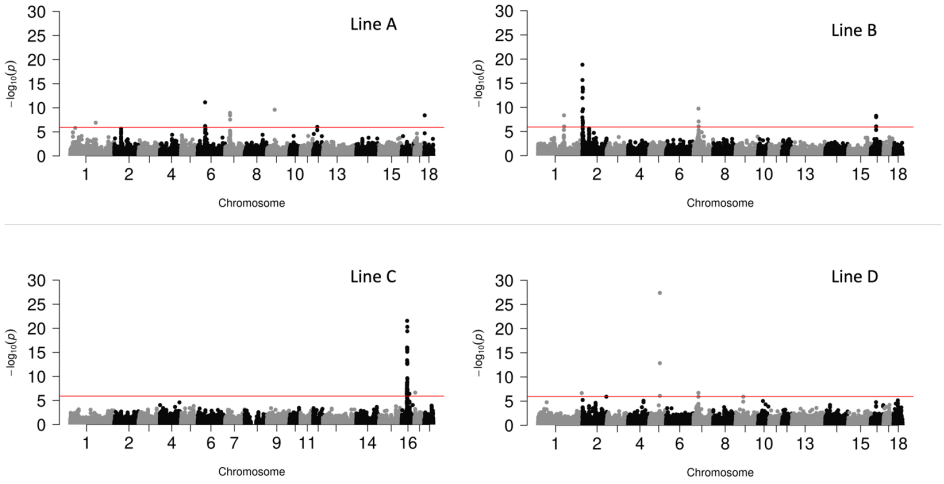

Figure 1. Genome wide plot of –log10 P values (y axis) for association of SNPs with loin depth in four pig lines. The horizontal line indicates the threshold for genome wide significance.

References

Bergamaschi, M., Maltecca, C., Fix, J., Schwab, C., & Tiezzi, F. (2020). J. Anim. Sci. 98(1), 1–11. doi.org/10.1093/jas/skz360

Bian, C., Prakapenka, D., Tan, C., Yang, R., Zhu, D. *et* al. (2021). Genet. Sel. Evol. 53(1). doi.org/10.1186/s12711-021-00661-y

Cánovas, A., Quintanilla, R., Amills, M., & Pena, R.N. (2010). In BMC Genom. (Vol. 11). http://www.biomedcentral.com/1471-2164/11/372

de las Heras-Saldana, S., Chung, K. Y., Kim, H., Lim, D., Gondro, C. *et al.* (2020). Genes, 11(11), 1–17. doi.org/10.3390/genes11111381

Derks, M. F. L., Gross, C., Lopes, M. S., Reinders, M. J. T., Bosse, M. *et al. (*2021). Genomics, 113(August 2020), 2229–2239.

Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Genet. Sel. Evol. 41(1), 55. doi.org/10.1186/1297-9686-41-55

Gozalo-Marcilla, M., Buntjer, J., Johnsson, M., Batista, L., Diez, F., *et al.* (2021). Genet. Sel. Evol. 53(76).

Groß, C., Derks, M., Megens, H.J., Bosse, M., Groenen, M.A.M. *et al.* (2020). Genet. Sel. Evol. 52(1). doi.org/10.1186/s12711-020-0528-9

Kim, K.S., Larsen, N., Short, T., Plastow, G., & Rothschild, M.F. (2000). Mamm. Genome. 11(2), 131–135. doi.org/10.1007/s003350010025

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. *et al.* (2014). Nat. Genet. 46(3), 310–315. doi.org/10.1038/ng.2892

Reyer, H., Varley, P. F., Murani, E., Ponsuksili, S., & Wimmers, K. (2017). Sci. Rep. 7(1). doi.org/10.1038/s41598-017-08961-4

Roudbari, Z., Coort, S. L., Kutmon, M., Eijssen, L., Melius, J., *et al.* (2020). Fron. Genet. 10. doi.org/10.3389/fgene.2019.01370

VanRaden, P.M., & Wiggans, G.R. (1991). J. Dairy Sci. 74(8), 2737–2746. doi.org/10.3168/jds.S0022-0302(91)78453-1

Zhou, S., Ding, R., Meng, F., Wang, X., Zhuang, Z., *et al.* (2021). BMC Genom. 22(1). doi.org/10.1186/s12864-020-07288-1

Zhou, X., & Stephens, M. (2012). Nat. Genet, 44, 821–824.

Zhuang, Z., Li, S., Ding, R., Yang, M., Zheng, E., Yang, H., *et al.* (2019). PLoS one., 14(6), 1–21. doi.org/10.1371/journal.pone.0218263