Accelerated discovery of functional genomic variation in pigs

Derks, Martijn F.L.; Groß, Christian ; Lopes, Marcos S. ; Reinders, Marcel .J.T.; Bosse, Mirte; Gjuvsland, Arne B. ; de Ridder, Dick; Megens, Hendrik-Jan; Groenen, Martien A.M.

DOI
10.1016/j.ygeno.2021.05.017

Publication date
2021

Document Version
Final published version

Published in
Genomics

Genomics 113 (2021) 2229–2239

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Accelerated discovery of functional genomic variation in pigs

Martijn F.L. Derks [a,d,*,1], Christian Gross [b,c,1], Marcos S. Lopes [d,e], Marcel J.T. Reinders [c], Mirte Bosse [a], Arne B. Gjuvsland [f], Dick de Ridder [b], Hendrik-Jan Megens [a], Martien A. M. Groenen [a]

[a] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, the Netherlands
[b] Bioinformatics Group, Wageningen University and Research, P.O. Box 633, 6708 PB Wageningen, the Netherlands
[c] Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands
[d] Topigs Norsvin Research Center, Beuningen, the Netherlands
[e] Topigs Norsvin, Curitiba, Brazil
[f] Norsvin SA, Hamar, Norway

ARTICLE INFO

Keywords:
Animal genomics
Functional variation
Genotype-phenotype link

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 number of genes contributing to the observed phenotype. Despite the vast increase of molecular data, pinpointing the causal variant underlying a phenotype of interest is still challenging. In this study, we present an approach to map causal variation and molecular pathways underlying important phenotypes in pigs. We prioritize variation by utilizing and integrating predicted variant impact scores (pCADD), functional genomic information, and associated phenotypes in other mammalian species. We demonstrate the efficacy of our approach by reporting known and novel causal variants, of which many affect non-coding sequences. Our approach allows the disentangling of the biology behind important phenotypes by accelerating the discovery of novel causal variants and molecular mechanisms affecting important phenotypes in pigs. This information on molecular mechanisms could be applicable in other mammalian species, including humans.

1. Background

Closing the gap between genotype and phenotype is a major goal in the life sciences, but remains extremely challenging [28]. Part of the complexity stems from phenotypes being influenced by many genes. Genome-wide association studies (GWAS) have been instrumental in statistically linking genotypes and phenotypes. These studies, resulting in identification of quantitative trait loci (QTL), have resulted in better understanding of the genomic architecture of complex traits [65]. However, the resolution of GWAS is limited since it only requires correlation to phenotypes by neighbouring markers in linkage disequilibrium (LD) Hence, unravelling the molecular drivers (causal variants) underlying phenotypes of interest requires further fine mapping [23]. The majority of causal variants are expected to reside in the noncoding regions of the genome, in particular in transcriptional regulatory regions [59], which can be very difficult to predict.

In human genetics, a combination of statistical fine-mapping methods and expression QTL (eQTL) studies are used to decrease the number of candidate genes and causal variants [8]. Further, functional annotation, facilitated by large consortium efforts including the Encyclopaedia of DNA Elements (ENCODE, [18]), can be applied to prioritize variants based on the likelihood of the variant(s) affecting gene expression. Despite this effort, identifying causal variants remains difficult, partly because of the fundamental complexity of phenotype-genotype relations, in which the environment also plays an important role.

In livestock, economically important phenotypes are typically determined by many genes, each explaining a small fraction of the phenotypic variation. However, for many traits it is now known that QTLs exist that explain more than 1% of the variation. For these

* Corresponding author at: Wageningen University & Research, Animal Breeding and Genomics, Wageningen, the Netherlands.
E-mail addresses: martijn.derks@wur.nl (M.F.L. Derks), marcos.lopes@topigsnorsvin.com (M.S. Lopes), M.J.T.Reinders@tudelft.nl (M.J.T. Reinders), mirte.bosse@wur.nl (M. Bosse), arne.gjuvsland@norsvin.no (A.B. Gjuvsland), dick.deridder@wur.nl (D. de Ridder), hendrik-jan.megens@wur.nl (H.-J. Megens), martien.groenen@wur.nl (M.A.M. Groenen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2021.05.017
Received 28 August 2020; Received in revised form 30 March 2021; Accepted 17 May 2021
Available online 20 May 2021

relatively large QTLs it is of interest to identify the underlying causal variation to better understand the biology of traits. Due to intense selection, the effective population size (N_e) of most livestock populations is small [33]. This often leads to a high degree of LD. This results in QTL comprising up to millions of base pairs (Mb) in length, especially in regions with low recombination rates [77]. 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. In contrast, livestock populations are less confounded by population stratification (i.e. ancestry differences between cases and controls), which can be a major problem in human GWAS studies [35].

Similar to human medical studies, further functional genomic information can help to prioritize the variants underlying the phenotypes of interest in livestock [63]. In pigs, the level of functional genome information, in contrast to human genome information, is limited. Fortunately, recent advances have been achieved in pigs by developing the pig Combined Annotation-Dependent Depletion (pCADD) tool [31], providing impact scores of all possible single-nucleotide substitutions in the pig genome. The CADD tool was originally developed to score variants with respect to their putative deleteriousness to prioritize potentially causal variants in human genetic studies [61]. This tool is frequently used to score variants in human GWAS studies [8]. Subsequently, other species-specific CADD tools were developed [30]. This tool scores the deleteriousness (or functional impact) of single nucleotide variants (SNPs) and it is built on a number of 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 $-10\log_{10}$ 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 and 10, the least 99% a score between 0 and 20 and so forth.

Pig populations have been under a long-term selection process performed by animal breeders to constantly improve their stock [41]. Therefore, such commercial breeds can be seen as a long-term controlled biological experiment. In general, genomic selection uses SNP chip variant panels to associate genomic regions with traits of interest. The variants of the panel are distributed across the genome and allows within-population genetic variation to be captured [50]. However, genomic selection uses the genome as a "black box", as the SNPs on the chip are mostly not causal, but genetically linked (by LD) to the actual causal variants and genes [32]. Therefore, the efficacy of genomic selection can be substantially improved by adding new genetic markers comprising the actual causal variation [29], providing insight into the exact molecular drivers involved in the phenotype.

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 demonstrate that pCADD scores can be used to identify causal variants underlying important phenotypes. Being able to identify causal variants will have major implications for genomic selection, and we show that CADD scores are promising in identifying causal variants in more neutral phenotypes as well. This study provides insights into the molecular biology and pathways affecting important phenotypes in pigs, that can also be transferred to human phenotypes.

2. Results

Here we present an approach to identify causal variants underlying important phenotypes in pigs. Our approach starts with a large scale GWAS study to identify loci associated with important phenotypes. Subsequently, we integrate the population whole genome sequence, the impact scores (pCADD), and further functional genomic information to fine map and report known and novel causal variants.

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

We analysed comprehensive genotype and phenotype data in four purebred pig populations: two boar breeds (Duroc and Synthetic), and two sow breeds (Landrace and Large White). In pigs, purebred populations are the units in which selection is applied, while the final production animals are derived from three-way crosses. First, F1 crossbred sows are created by mating two sow breeds selected for high reproductivity and mothering ability, which are subsequently crossed with a boar breed especially selected for meat production traits. The examined traits can be grouped in three classes: (1) traits focussing on production traits, including backfat, intramuscular fat, growth rate and feed efficiency; (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, congenital defects, 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 271 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 was used as a starting point for further analysis.

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

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

In our approach the first step entails identification of the SNP of highest significance in a GWAS peak. Subsequently, all variants that are in high LD ($r^2 > 0.7$) were extracted. The variants were 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, all high-LD variants were annotated for their pCADD scores, to prioritize them for likely impact on phenotype. The sequence variants were also annotated for variant effect type, using the Ensembl Variant Effect Predictor (VEP, release 98) [49]. 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 Supplementary Table 1. pCADD provides an independent impact score and both, deleterious and functional SNPs, will be enriched for high pCADD scores, because they have impact (either negative or positive). The assumption is that if a variant has impact on a trait (either regulatory or coding), it likely falls within a rather evolutionary conserved region, leading to generally higher pCADD scores. 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 [78].

2.2.2. 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 [74]. 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 humans, and gene knockouts in mice and rats [83]. A complete overview of the pipeline is presented in Fig. 1.

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

The combination of genotype and gene expression data provides an

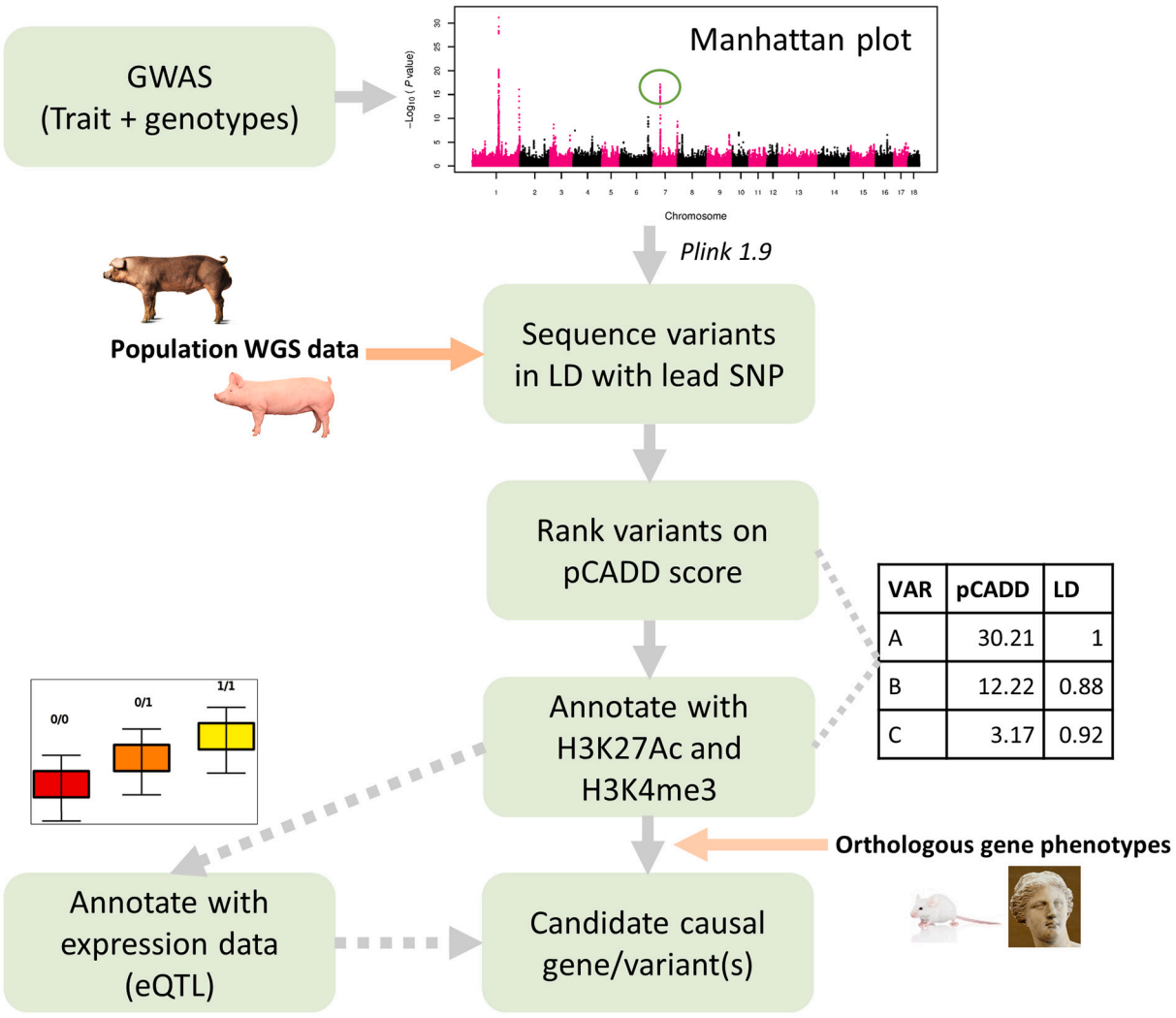

Fig. 1. Pipeline overview. The pipeline takes the result of a GWAS as input (lead SNP) and identifies SNPs from WGS data that are in high LD with the lead SNP. Subsequently, the variants are prioritizes based on impact scores (pCADD), open chromatin information (liver), and gene expression (if available).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

additional layer of evidence to find causal variants, as differences in expression of genes can be associated with a specific variant (expression quantitative trait loci; eQTL). In this study we use 59 RNA-sequenced samples [75] 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 sequenced RNA samples were derived from testis. Further details about the sequenced samples and alignment depth are provided in Supplementary Table 2. The combination of epigenomic marks (liver) and gene-expression data (testis) can, in addition to the pCADD scores, facilitate the discovery of functional variants.

2.3. Accelerated discovery of potential causal variants from GWAS results

To demonstrate the power of our approach we first analysed several QTL regions (per breed) with known causal variants reported in literature. This list includes 1) a missense mutation in *MC4R* affecting production traits [39], 2) a variant in the promoter of the *VRTN* gene affecting number of teats [76], and 3) a missense mutation in *PRKAG3* affecting meat quality [51]. Despite the fact that hundreds of variants were found to be in LD at each of the GWAS peaks, the method returned the causal variant as top ranked for both the *MC4R* missense mutation (Supplementary Text 1, Supplementary Fig. 1) and the *VRTN* promoter variant (Supplementary Text 2, Supplementary Fig. 2, Supplementary Fig. 3, Supplementary Table 3).

The mutation in the *PRKAG3* gene identified by Milan et al. [51] does not segregate in our sequenced animals. However, we identified another missense variant (15:g.120865869C > T) in the *PRKAG3* gene that is a strong candidate variant for affecting meat quality in both boar breeds (Fig. 2), as described by [73]. The causal missense variant is highlighted in green, and the lead SNP in the GWAS in blue in Fig. 2b. The variant results in a substitution of glutamic acid for lysine (ENSSSCP00000030896:p.Glu47Lys) (Supplementary Table 4). *PRKAG3* regulates several intracellular pathways, including glycogen storage [19]. 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 [48]. Gain of function mutations in the *PRKAG3* gene have been correlated with increased glycogen content in skeletal muscle in pig, negatively affecting meat quality [15]. 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.

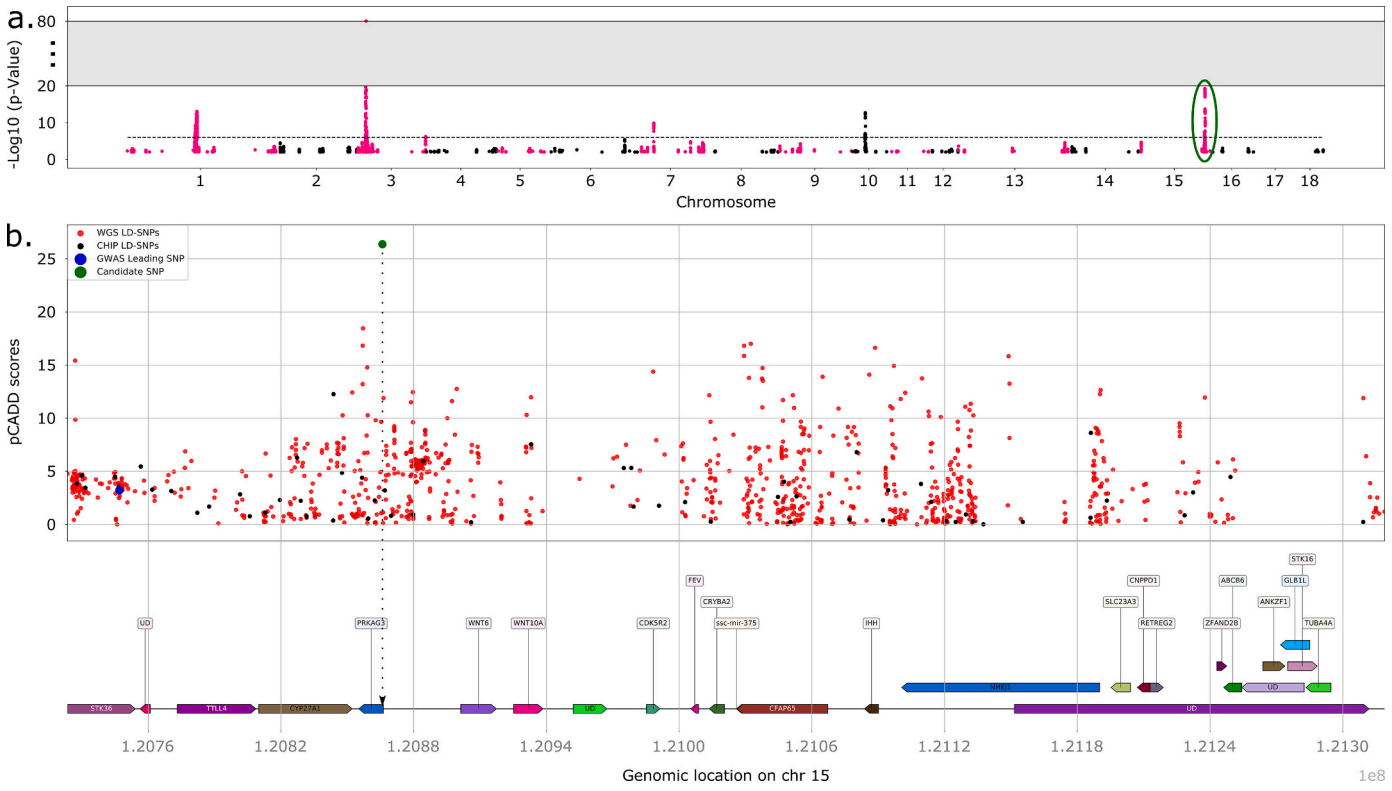

Fig. 2. a) Manhattan plot for drip loss in Duroc showing a strong QTL on chromosome 15:121 Mb. Only SNPs with a -$\log_{10}$(p) > 2 are plotted. Magenta and black dots distinguish between chromosomes. b) Plot showing all sequence variants in high LD (red) with the lead SNP (blue) from the QTL region highlighted in the green circle in panel A, 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.4. Integrated analysis reveals several novel variants with pleiotropic effects on important phenotypes

We systematically analysed all significant QTL regions (−log10(P) > 6.0) to identify novel causal variants affecting important phenotypes. We highlight several of the most striking variants in the subsequent paragraphs, and a summary of likely causal loci is given in Table 1. In addition, we show several examples of highly significant QTL regions for which we were unable to assign a candidate causal variant at the end of the results.

2.4.1. 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 (Fig. 3a). This QTL is among the most significant region affecting production traits (together with the *MC4R* QTL on chromosome 1 and a strong QTL on chromosome 10). The lead SNP in the GWAS result is located at position 7:30,116,227 with a -$\log_{10}$(p) > 20 for backfat, feed consumption, and intramuscular fat (Supplementary Fig. 4). The analysis returned 485 variants in high LD with the lead SNP (Fig. 3b). The two variants with the highest pCADD scores are upstream of the *HMGA1* gene, 566 bp apart (Fig. 3b, Supplementary Table 5). Both mutations are in the promoter region of the *HMGA1* gene. A promotor function at this position is supported by signals on the H3K4me3 and H3K27Ac histone marks (Supplementary Fig. 5). The A allele, segregating at 36% allele frequency, is associated with less backfat, faster growth, but also smaller loin depth and decreased intramuscular fat. The expression of the *HMGA1* gene was investigated in twenty samples. In all samples genotype and gene expression, as normalized fragments per kilobase per million (FPKM), were both available within the three genotype classes GG, AG, and AA. The A allele is strongly associated with increased expression of the gene. The expression level furthermore appears to be affected in an additive manner ($P = 0.041$, Supplementary Fig. 6). The correlation of additive increase of expression of the *HMGA1* gene, and the same additive effect on backfat and growth, and negative on intramuscular fat, suggests a causal relation between gene expression and phenotype. In addition, we find two variants affecting the promoter region of the *HMGA2* gene, associated with less backfat in the Synthetic breed (Table 1, Supplementary Table 6). Both *HMGA1* and *HMGA2*, belong to the High Mobility Group A gene family, are well-known to affect growth and stature in pigs [14,36,40], although a causal relationship had not been reported thus far. Our results suggest that the causal variants for both genes are regulatory.

2.4.2. 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 (Fig. 4a). Despite the presence of more significant QTL regions we focus on this QTL given the large extend of LD. Large LD blocks can hamper accurate fine mapping of causal variants. The lead SNP in the GWAS result is located at position 1:115,884,118. The analysis returned 874 variants in high LD with the lead SNP (Supplementary Table 7). The SNP with the highest pCADD score (1:g.120074006G > A, pCADD = 30.28), a single missense variant affecting the *SCG3* gene, was identified as potentially causal (Fig. 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 altering the *SCG3* protein have been associated with obesity in humans [69], supporting its likely causality for the fat-associated phenotypes in pigs.

Table 1
List of potential causal variants identified from the pipeline. Table shows the variants type, potential overlap with promoter or enhancer region (from liver, [78]), 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 has already been reported in literature. A minus sign stands for a negative effect of the alternative allele in the table (orange), a plus sign stands for the positive effect of the alternative allele on the indicated trait (green). NS indicates that the variant has no significant effect on the trait. Variant IDs are given in Supplementary table 16. [1,4,10,16,20,21,45,53,54,56,62,67,72,82]

Chr	Variant	Type	Promotor/ Enhancer	Amino acid change	pCADD score	Rank	Gene	Breed(s)	BFE	IMF	TGR	DRY	NTE	Supporting evidence
1	G-120074006-A	missense	NO	T386M	30.27	1	*SCG3*	Synthetic	-	+	NS	+	NS	Associated with obesity (Tanabe et al. 2007).
1	G-160773437-A	missense	NO	D298N	27.47	1	***MC4R***	Synthetic, Duroc	-	NS	+	NS	NS	Associated with fatness, growth, and feed intake traits (Kim et al. 2000).
7	G-30318881-A	upstream	YES	-	14.96	1	*HMGA1*	Duroc	-	+	+	+	NS	Associated with pig growth and fat deposition traits (Kim et al. 2006).
5	T-30187091-C	upstream	NO	-	19.44	2	*HMGA2*	Synthetic	-	NS	NS	NS	NS	*HMGA2* deficiency in pigs leads to dwarfism (Chung et al. 2018).
18	T-10098558-C	intron	NO	-	16.41	1	*HIPK2*	Synthetic	-	-	NS	NS	NS	Essential regulator of white fat development (Sjolund et al. 2014).
2	A-144841051-C	intron	NO	-	10.65	2	*NR3C1*	Synthetic, Large White	+	+	NS	NS	NS	Glucocorticoid receptor is a transcription factor activated by circulating glucocorticoids and mediates their effects on various biological functions in the body (Reyer et al. 2013).
5	G-65814519-A	missense	NO	V850I	23.10	1	*AKAP3*	Duroc	+	NS	NS	NS	NS	Candidate gene for meat tenderness traits (Casiro et al. 2017).
11	T-20619202-C	3'UTR	NO	-	18.46	1	*HTR2A*	Duroc	-	NS	NS	NS	NS	Positive role of *HTR2A* in adipogenesis (Yun et al. 2018).
13	A-195332161-G	intron	YES	-	6.13	26	*SOD1*	Duroc	-	NS	NS	NS	NS	Associated with abnormal body fat mass (Liu et al. 2013).
3	C-94863278-A	5'UTR	YES	-	16.11	7	*PRKCE*	Landrace	-	NS	NS	NS	NS	Associated with decreased total body fat amount (Castrillo et al. 2001).
14	G-128748846-A	5'UTR	YES	-	15.53	7	*CACUL1*	Synthetic	NS	-	NS	NS	NS	Exhibits a repressive role in PPARγ activation and fat accumulation (Jang et al. 2017).
2	T-103610859-C	missense	NO	L335S	21.45	1	*LNPEP*	Synthetic	NS	+	NS	NS	NS	Decreased white fat cell size, decreased susceptibility to diet-induced obesity (Niwa et al. 2015).
2	C-41019232-T	upstream	NO	-	3.91	9	*SAA3*	Large White	NS	+	NS	NS	NS	Decreased susceptibility to diet-induced obesity, increased white fat cell size (den Hartigh et al. 2014).
4	A-88412353-C	intron	NO	-	18.99	1	*NOS1AP*	Large White	NS	+	NS	NS	NS	Regulates glucose homeostasis and hepatic insulin sensitivity in obese mice (Mu et al. 2019).
15	C-117292901-A	missense	NO	G1693C	24.75	1	*ABCA12*	Large White	NS	+	NS	NS	NS	Associated with pig production traits (Piorkowska et al. 2014).
6	A-146830209-G	intron	NO	-	11.88	1	*LEPR*	Duroc	NS	+	NS	NS	NS	The gene expression results showed that in the loin muscle LEPR showed significantly higher expression in pigs with higher IMF% (Li et al. 2010).
7	A-11391274-G	intron	NO	-	9.72	1	*JARID2*	Duroc	NS	+	NS	NS	NS	Regulates cell-cycle in skeletal muscle (Adhikari et al. 2019).
2	G-15310202-A	3' UTR / missense	NO	-	21.38	1	*NR1H3 / MADD*	Synthetic	NS	NS	-	NS	NS	*NR1H3*: Obesity, associated with lipid deposition in pigs (Zhang et al. 2016).
5	A-83681067-G	intron	YES	-	12.10	9	*NR1H4*	Synthetic	NS	NS	-	NS	NS	Glucose tolerance, lipid homeostasis (Sinal et al. 2000).
9	C-17077403-A	5' UTR	YES	-	8.45	7	*PRCP*	Synthetic	NS	NS	-	NS	NS	Reduced levels of PRCP promote obesity (Palmiter 2009).
12	C-44684331-G	missense	NO	G131R	25.81	1	*SLC46A1*	Synthetic	NS	NS	+	NS	NS	Decreased total body fat amount, decreased circulating glucose levels (Blake et al. 2017).
6	A-145977262-T	intron	NO	-	14.63	1	*SGIP1*	Large White	NS	NS	-	NS	NS	Suppression of SGIP1 reduced body weight (Trevaskis et al. 2005).
15	C-120865869-T	missense	NO	E47K	26.37	1	***PRKAG3***	Synthetic, Duroc	NS	NS	NS	-	NS	A combination of two variants in *PRKAG3* is needed for a positive effect on meat quality in pigs. (Uimari and Sironen 2014)
6	C-67433001-T	intron	YES	-	11.87	2	*KLHL21*	Landrace	NS	NS	NS	+	NS	Affects creatinine levels in mice (Blake et al. 2017).
1	C-127921686-T	missense	NO	G1904S	23.03	1	*MAP1A*	Synthetic	NS	NS	NS	-	NS	Involved in glycogen synthesis (Halpain and Dehmelt 2006).
2	C-96202720-T	intron	NO	-	17.86	2	*MEF2C*	Synthetic	NS	NS	NS	-	NS	*MEF2C* skeletal muscle knockout mice accumulate glycogen in their muscle (Anderson et al. 2015).
9	C-9329652-T	missense	NO	P419S	20.91	3	*NEU3*	Synthetic	NS	NS	NS	-	NS	Overexpression increases glycogen deposition (Yoshizumi et al. 2007).
13	G-173634576-A	upstream	YES	-	15.68	1	*GBE1*	Synthetic	NS	NS	NS	+	NS	Glycogen branching enzyme (Froese et al. 2015).
9	G-758928-A	missense	NO	A768T	21.92	1	*TRIM66*	Synthetic	NS	NS	NS	-	NS	*TRIM66* is involved in regulating glycogen synthesis (Fan et al. 2019).
14	T-107058908-C	intron	YES	-	24.50	1	*SORBS1*	Synthetic	NS	NS	NS	-	NS	Glycogen binding protein (Nagy et al. 2018).
15	A-46758359-G	intron	YES	-	11.73	4	*SORBS2*	Synthetic	NS	NS	NS	-	NS	-
7	A-97614602-C	upstream	NO	-	11.95	2	***VRTN***	Duroc, Landrace, Large White	NS	NS	NS	NS	+	Associated with increased number of vertebrae in pigs (van Son et al. 2019).
8	A-102781174-G	missense	NO	M165V	21.27	1	*QRFPR*	Synthetic	NS	NS	NS	NS	-	The G-protein-coupled receptor *QRFPR* regulates bone formation (Baribault et al. 2006).

2.5. Integrated analyses reveal the molecular pathways involved in selection on production traits

2.5.1. Genes affecting meat quality involved in muscle glycogen storage

A substantial number of candidate causal variants affecting meat quality were identified. In the Synthetic breed, a boar line hence particularly selected for its meat quality traits, 26 loci were significantly associated with drip loss ($-\log_{10}(p) > 6$). Drip loss is trait that measures the water holding capacity of the meat (Fig. 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 are known to lead to increased drip loss, which is considered to negatively affect meat quality [64]. Examples discovered in the present study include regulatory variants affecting the *MEF2C, SCG3,* and *GBE1* genes. *MEF2C* knockout mice accumulate glycogen in their muscles [2], while *GBE1* codes for a glycogen branching enzyme associated with glycogen storage disease, if mutated [22]. 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 [34,81]. Not only does our study demonstrate the central role of glycogen-based pathways in an important meat quality trait, it also highlights that a combination of regulatory and protein altering variants are involved.

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

A number of likely causal variants and genes affecting other important production traits were found (Table 1), although the underlying pathways initially appeared to be less obvious. The top-ranked genes were found to be enriched in energy reserve metabolic processes, glycogen metabolic process, regulation of lipid biosynthetic process, and homeostasis (Supplementary Table 8). More specifically, two regulatory variants in the *SOD1* and *PRKCE* genes likely affect backfat. *SOD1* is involved in glucose metabolism and prevents oxidative damage associated with obesity in humans [46], while mutations in *PRKCE* decrease the amount of body fat in humans [11]. Furthermore, we identified one regulatory variant in the *CACUL1* gene likely affecting intramuscular fat. This gene inhibits adipogenesis via the peroxisome proliferator-activated receptor γ (PPARγ) [38]. In addition, two missense variants likely 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 [55], while the *ABCA12* gene plays an important role in lipid transport, affecting carcass fat content in pigs [58]. We further identified potential regulatory variants in the *NR1H3*, *NR1H4*, and *PRCP* genes, all likely affecting growth (Table 1). Note that the 3'UTR variant with the highest pCADD score in the *NR1H3* gene is also causing a missense variant in the partly overlapping *MADD* gene. *NR1H3* and *NR1H4* are paralogous genes both involved in lipid homeostasis [66,84], 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 [7]. The various pathways identified to be involved in the physiology of growth and metabolism demonstrate that indeed selection traits involved in growth, energy efficiency, and fat deposition are complex and consisting of many genes, which is congruent with the known highly quantitative nature of these traits. However, identifying these underlying pathways enable identifying at which genes and parts of gene networks these pathways intersect, which provides valuable insight into pleiotropy and their trade-

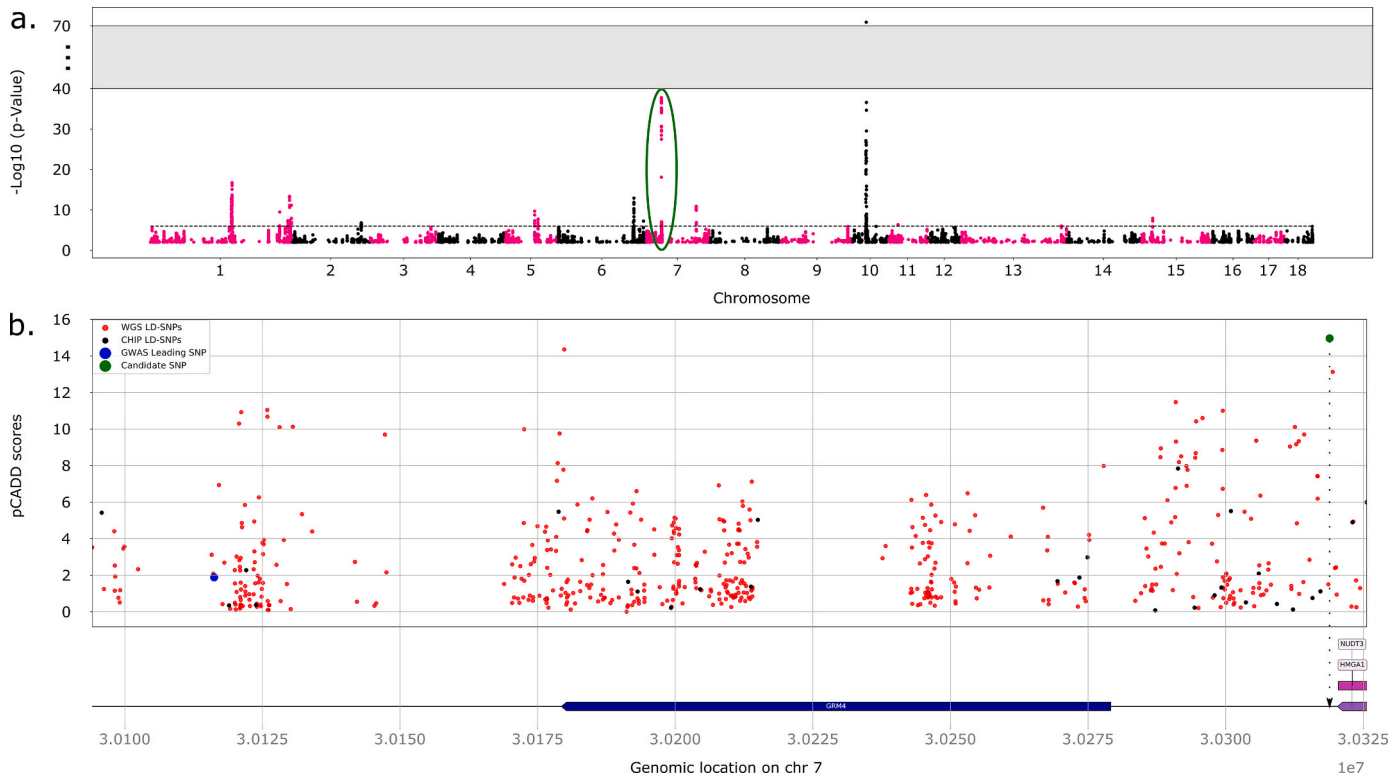

Fig. 3. a) Manhattan plot for backfat in Duroc showing a strong QTL on chromosome 7:30 Mb. Only SNPs with a -$\log_{10}$(p) > 2 are plotted. Magenta and black dots distinguish between chromosomes. b) Plot showing all sequence variants in high LD (red) with the lead SNP (blue) from the QTL region highlighted in the green circle in panel A, 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

offs.

2.5.3. *Balancing selection for causal variants in the breeding program*

Interestingly, a number of the variants that are likely directly affecting phenotypes under selection in commercial breeding programs, exhibit pleiotropic effects. This particularly applies to genes *HMGA1*, *SCG3*, 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 backfat. 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. Without balancing selection, causal variants of this magnitude of impact would have been selected to fixation in modern breeding programs in a very low number of generations.

2.6. *Examination of three highly significant across-breed QTL regions*

We examined three other striking QTL regions (from Figs. 2–4) to identify potentially causal mutations. First, we examined the QTL region for drip loss on chromosome 3 (Supplementary Table S12). The lead SNP (3:16839270) is in high LD ($r^2 = 0.92$) with the splice variant known to affect the expression of the *PRHG1* gene affecting meat quality [47]. However, the splice variant is not highly scored according to pCADD (score = 2.13) and is thereby not among the top variants. Next, we examined the QTL region on chromosome 10 affecting the production traits backfat, feed intake, and intramuscular fat (Fig. 3, Supplementary Fig. 4). The top 5 candidate variants affect the *ZNF367* (missense variant), *CDC14b* (intronic), and *AAED1* (splice donor) genes (Supplementary Table S14). The *ZNF367* increases creatine levels in knockout mice [7]. The *CDC14b* gene is involved in DNA damage repair and aging [12], while the *AAED1* gene is involved in glycolysis [85]. Further analysis is required to fine map the causal variant at this locus. Next, we examined the QTL region for backfat on chromosome 5 in the Synthetic breed (Fig. 4), the top SNPs are intergenic or annotated within the intronic regions of the *CCND2* gene (Supplementary Table S13). *CCND2* is involved in the cell cycle and has been reported to affect fat deposition traits in pigs [42].

3. Discussion

The aim of this study was to identify causal variants under selection in pig breeding programs, and to identify the molecular pathways involved in the traits. Including the pCADD scores is particularly relevant because genomic variation underlying phenotypic variation mostly affects the non-coding part of the genome [59], and GWAS results often point to regions outside gene boundaries [5]. Furthermore, the extensive LD in regions under selection makes it hard to pinpoint any single variant since there may be several candidates that all may be significant in a GWAS. pCADD scores [31] allow the prioritization of any single nucleotide 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 *epi*genomics 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 [61]. However, the pCADD scores are built on far less (epi)-genomics data, but with the accumulation of functional genomic data in pigs [27], these pCADD scores will further improve. One drawback of pCADD is that it is not able to score structural variants yet. Advances have been made for the human pCADD to include scoring of structural variants [24] and this feature might be added in

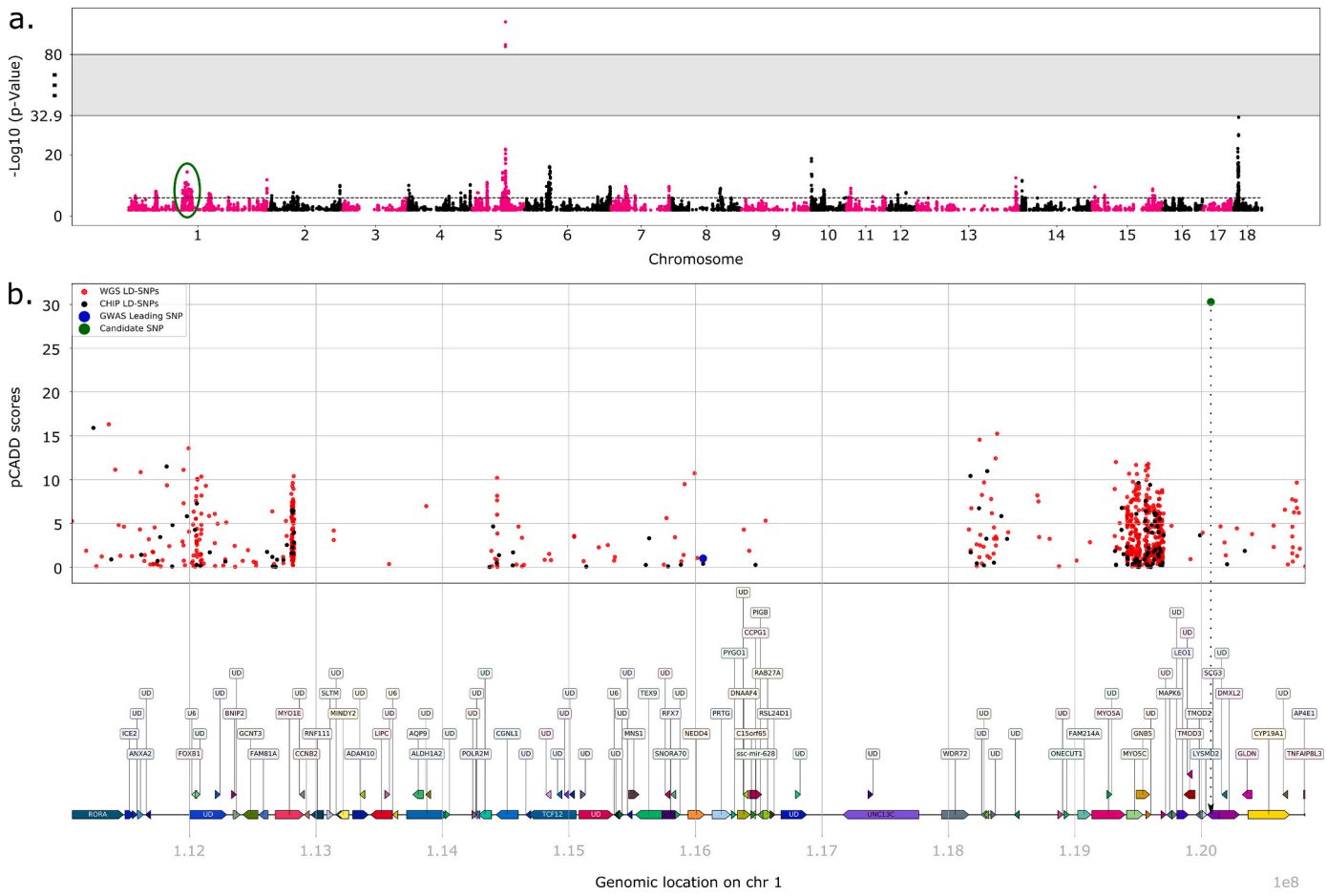

Fig. 4. a) Manhattan plot for backfat in the Synthetic breed showing a strong QTL on chromosome 1:116 Mb. Only SNPs with a $-\log_{10}(p) > 2$ are plotted. Magenta and black dots distinguish between chromosomes. b) Plot showing all sequence variants in high LD (red) with the lead SNP (blue) from the QTL region highlighted in the green circle in panel A, 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

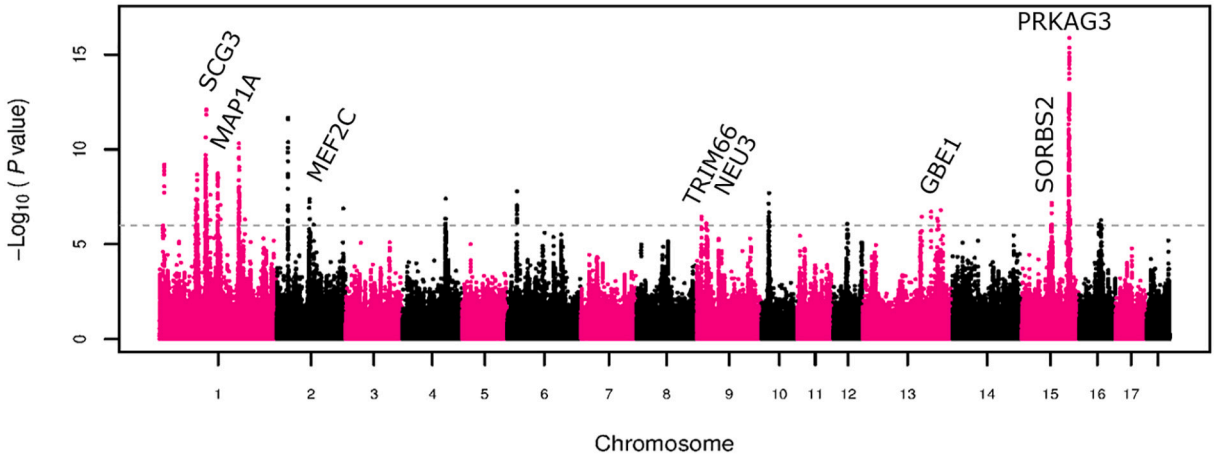

Fig. 5. Manhattan plot for drip loss in the synthetic breed. The figure shows significant loci and likely causal genes identified.

future updates of pCADD.

Livestock populations generally have small effective population sizes (N_e: 50–200), far less compared to e.g. human ($N_e \sim 10{,}000$), 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 (Supplementary Fig. 3). 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. GWAS analysis are generally performed within breed. However, a multibreed GWAS (potentially including crossbred information) could be performed to disentangle the haplotype structures in different breeds, and to facilitate fine mapping of the causal variant. Hence, future GWAS analysis and finemapping strategies could benefit from a multi-breed approach. Another way to provide further evidence of causality would be to fit the SNP into the GWAS model and check whether the QTL peak disappears completely. This would however require imputation to sequence and will only benefit if the LD between the causal SNP and the lead SNP in the 660 K GWAS is significantly below 1.0. This example shows that integrating GWAS and pCADD scores can be very powerful to prioritize variants. There is, however, a trade-off: although the homogenous populations lend themselves well for finding associations by GWAS, the extensive linkage disequilibrium results in too many high-pCADD SNP candidates. But this does also highlight a second aspect that is often not considered: that strong selection in small, homogeneous populations may lead to strong hitchhiking effects. The type of analysis presented in this study provides a strong method for inferring potential hitchhiking and inbreeding effects, given the knowledge on the functional variant and the surrounding (possible deleterious) variants in high LD. Note that in pig breeding the production animals are crossbreds not affected by inbreeding that benefit from the heterosis effect. However, we predict that it will become very valuable to genotype causal variants in crossbred animals, especially because the LD between the selection markers and the causal variants might be substantially lower in crossbred animals. Hence, we believe that causal variants could significantly improve across-breed genomic prediction.

Although the development of genomic selection has revolutionized animal breeding, the lack of functional genomic information currently limits further development [26]. The framework and associated pCADD scores provided within this study will accelerate the discovery of functional variants, which can be directly implemented in genomic selection by adding the causal variants to the SNP chips used for genomics selection. Moreover, the results provide further knowledge of the biological pathways associated with important phenotypic variation in livestock. This is vitally important, since breeding goals are in practice often mutually exclusive. Understanding how at a fundamental biological level, pathways under selection are intersecting, can provide a better formulation of selection criteria.

Integrating GWAS based on ongoing commercial selection and functional appraisal of variations in the populations under selection provides a powerful framework to study the genetic architecture of the traits under selection. Comparing the pathways and genes found to be important in these traits in this manner, reveals a striking functional overlap in similar phenotypes in other mammals. 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 [3]. 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 uniform conditions, which could make the pig an ideal model to study the effects of specific genic variants on these analogous phenotypes [57]. Pig breeding has led to extreme changes in animal production and efficiency, with little negative consequences on health [41]. This remarkable robustness of the animals, and the molecular mechanisms involved, may aid in understanding metabolic disease in human.

Ultimately, quantitative selection seeks to perturb the underlying pathways in commercial traits. Our study suggests that, despite the complexity of pathways and the high number of genes potentially involved in any one trait, there may be a small number of genes that are exceptionally suited as 'entry points' into those pathways. These genes have a large effect that are more likely to be under selection than other genes in the same pathway. Understanding these 'key' genes, and how they function together would further help to unravel the (molecular) consequences of selection.

4. 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 phenotypes. 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 humans and pigs.

5. Methods

5.1. 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).

5.2. Genotype data and breeds

The genomic dataset consists of 15,791 (Duroc), 28,684 (Synthetic), 36,956 (Large White), and 41,865 (Landrace) animals genotyped on the (Illumina) Geneseek custom 50 K SNP chip with 50,689 SNPs (50 K) (Lincoln, NE, USA) and imputed to the Axiom porcine 660 K array from Affymetrix (Affymetrix Inc., Santa Clara, CA, United States). The chromosomal positions were determined based on the Sscrofa11.1 reference assembly [79]. 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 0.90, and an animal call rate > 0.90. SNPs with a *p*-value below 1×10^{-12} for the Hardy-Weinberg equilibrium exact test were also discarded. All pre-processing steps were performed using Plink v1.90b3 [60].

5.3. Phenotypes

A total of 1,360,453 animals with phenotypic records for at least one of the 83 evaluated traits were available for this study. These animals were either purebred (Duroc, Synthetic, Landrace and Large White) or crossbred originated from the crosses between these purebred populations. The phenotypic records were used in the estimation of breeding values for all evaluated traits. The estimated breeding value (EBV) of each animal was obtained from the routine genetic evaluation by Topigs Norsvin applying the single-step approach [13,52], which allows the simultaneous evaluation of genotyped and non-genotyped animals, using the software MiXBLUP [70].

5.4. Genome wide association study

A single SNP GWAS was performed with the software GCTA [80] by

applying the following model:

$$EBV_j = \mu + SNP_i + a_j + e_{ij}$$

where EBV_j is the EBV of the genotyped animal j, μ is the overall EBV mean of the genotyped animals, 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. Association results were considered significant if -log10(p) > 6.0.

5.5. 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 sequence depth ranges from 6.6 to 22.2, with an average depth of 11.82 (Supplementary Table 10). FastQC was used to evaluate read quality [6]. BWA-MEM (version 0.7.15, [43]) was used to map the WGS data to the Sscrofa11.1 reference genome. SAMBLASTER was used to discard PCR duplicates [21], and samtools was used to merge, sort, and index BAM alignment files [44].

5.6. 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 [25]. Post processing was performed using BCFtools [44]. 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 and 2.35 (Supplementary Table 10). Variant (SNPs, Indels) annotation was performed using the Variant Effect Predictor (VEP, release 97) [49].

5.7. pCADD scores

pCADD scores were retrieved from Gross et al. [31] [31]. Visualization of pCADD scores was performed using JBrowse 1.16.6 [68]. Integration of sequence variants with pCADD score was performed using PyVCF [9]. 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/.

5.8. 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 [78] [78]. Data was aligned using BWA-mem [43] and visualized in JBrowse [68]. Coverage information on variant sites was obtained using PyVCF [9] and the PySAM 0.15.0 package.

5.9. Phenotypes and gene ontology

Phenotype information from genes orthologous to pigs in humans, mice, and rats were retrieved from the Ensembl database ([37], release 97) using a custom bash script. Gene ontology and pathway information was obtained from the UniProt database [74].

5.10. RNA-sequencing and differential expression

We used 25 Duroc and 34 Landrace 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 [75]. 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 [75]. 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 [17] and called transcripts and normalized FPKM expression levels using Cufflinks and Cuffnorm [71]. We assigned the genotype class (homozygous reference, heterozygous, homozygous alternative) for each RNA-sequenced individual using the 660 K 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.

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. In addition, this study was supported by the IMAGE project (Horizon 2020, No. 677353). Mirte Bosse was financially supported by NWO-VENI grant no. 016.Veni.181.050. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The use of the HPC cluster was made possible by CATAgroFood (Shared Research Facilities Wageningen UR). The data and some of the analyses used in this study was partly financed by the Research Council of Norway through the projects “Precision whole genome sequence to precision breeding” – NFR no. 255297/E50 and “Investigation of boar fertility by genetic characterization and detection of traits important in sperm production and quality” - NFR no. 207568/O99.

Author contributions

MD conceived the study. MD and CG planned the analysis framework, MD developed the analysis pipeline, CG provided scripts for the graphical output and to annotate candidate genes with phenotypes of orthologous genes in human, mouse and rat. MD took the lead in writing the manuscript and was supported by CG. ML conducted the GWAS and provided data. HJM, AG and MG contributed to the interpretation of the results. MR, MB and DdR helped to supervise the study. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Declaration of Competing Interest

ML and AG are employees of Topigs Norsvin, a research institute closely related to one of the funders (Topigs Norsvin). All authors declare that the results are presented in full and as such present no conflict of interest. The other Breed4Food partners Cobb Europe, CRV, Hendrix Genetics, declare to have no competing interests for this study.

Acknowledgments

The authors thank Egbert Knol and Barbara Harlizius for useful input on this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.05.017.

References

[1] A. Adhikari, P. Mainali, J.K. Davie, JARID2 and the PRC2 complex regulates the cell cycle in skeletal muscle, J. Biol. Chem. (2019), https://doi.org/10.1074/jbc.RA119.010060.
[2] [1].
[3] Y. Bao, P. Kishnani, J.Y. Wu, 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. Investig. 97 (4) (1996) 941–948.
[4] H. Baribault, J. Danao, J. Gupte, L. Yang, B. Sun, W. Richards, H. Tian, The G-protein-coupled receptor GPR103 regulates bone formation, Mol. Cell. Biol. 26 (2006) 709–717.
[5] N. Bartonicek, M.B. Clark, X.C. Quek, J.R. Torpy, A.L. Pritchard, J.L.V. Maag, B. S. Gloss, J. Crawford, R.J. Taft, N.K. Hayward, et al., Intergenic disease-associated regions are abundant in novel transcripts, Genome Biol. 18 (2017) 241.
[6] B. Bioinformatics, FastQC: A Quality Control Tool for High Throughput Sequence Data, Babraham Institute, Cambridge, UK, 2011.
[7] J.A. Blake, J.T. Eppig, J.A. Kadin, J.E. Richardson, C.L. Smith, C.J. Bult, the Mouse Genome Database G, Mouse genome database (MGD)-2017: community knowledge resource for the laboratory mouse, Nucleic Acids Res. 45 (2017) D723–D729.
[8] M.E. Cannon, K.L. Mohlke, Deciphering the emerging complexities of molecular mechanisms at GWAS loci, Am. J. Hum. Genet. 103 (2018) 637–653.
[9] J. Casbon, PyVCF - a Variant Call Format Parser for Python, 2012.
[10] S. Casiro, D. Velez-Irizarry, C.W. Ernst, N.E. Raney, R.O. Bates, M.G. Charles, J. P. Steibel, Genome-wide association study in an F2 Duroc x Pietrain resource population for economically important meat quality and carcass traits, J. Anim. Sci. 95 (2017) 545–558.
[11] A. Castrillo, D.J. Pennington, F. Otto, P.J. Parker, M.J. Owen, L. Bosca, Protein kinase C epsilon is required for macrophage activation and defense against bacterial infection, J. Exp. Med. 194 (2001) 1231–1242.
[12] M. Chiesa, M. Guillamot, M.J. Bueno, M. Malumbres, The Cdc14B phosphatase displays oncogenic activity mediated by the Ras-Mek signaling pathway, Cell Cycle 10 (2011) 1607–1617.
[13] O.F. Christensen, M.S. Lund, Genomic prediction when some animals are not genotyped, Genet. Sel. Evol. 42 (2010) 2.
[14] J. Chung, X. Zhang, B. Collins, R.B. Sper, K. Gleason, S. Simpson, S. Koh, J. Sommer, W.L. Flowers, R.M. Petters, 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. 115 (2018) 5420–5425.
[15] D. Ciobanu, J. Bastiaansen, M. Malek, J. Helm, J. Woollard, G. Plastow, M. Rothschild, Evidence for new alleles in the protein kinase adenosine monophosphate-activated gamma(3)-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality, Genetics 159 (2001) 1151–1162.
[16] L.J. den Hartigh, S.R. Wang, L. Goodspeed, Y.L. Ding, M. Averill, S. Subramanian, T. Wietecha, K.D. O'Brien, A. Chait, Deletion of serum amyloid A3 improves high fat high sucrose diet-induced adipose tissue inflammation and hyperlipidemia in female mice, PLoS One (2014) 9.
[17] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21.
[18] ENCODE, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[19] B. Essen-Gustavsson, A. Granlund, B. Benziane, M. Jensen-Waern, A.V. Chibalin, Muscle glycogen resynthesis, signalling and metabolic responses following acute exercise in exercise-trained pigs carrying the PRKAG3 mutation, Exp. Physiol. 96 (2011) 927–937.
[20] W. Fan, F. Du, X. Liu, TRIM66 confers tumorigenicity of hepatocellular carcinoma cells by regulating GSK-3beta-dependent Wnt/beta-catenin signaling, Eur. J. Pharmacol. 850 (2019) 109–117.
[21] G.G. Faust, I.M. Hall, SAMBLASTER: fast duplicate marking and structural variant read extraction, Bioinformatics 30 (2014) 2503–2505.
[22] D.S. Froese, A. Michaeli, T.J. McCorvie, T. Krojer, M. Sasi, E. Melaev, A. Goldblum, M. Zatsepin, A. Lossos, R. Alvarez, et al., Structural basis of glycogen branching enzyme deficiency and pharmacologic rescue by rational peptide design, Hum. Mol. Genet. 24 (2015) 5667–5676.
[23] M.D. Gallagher, A.S. Chen-Plotkin, The post-GWAS era: from association to function, Am. J. Hum. Genet. 102 (2018) 717–730.
[24] L. Ganel, H.J. Abel, C. FinMetSeq, I.M. Hall, SVScore: an impact prediction tool for structural variation, Bioinformatics 33 (2017) 1083–1085.
[25] E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, arXiv preprint (2012) arXiv:12073907.
[26] M. Georges, C. Charlier, B. Hayes, Harnessing genomic information for livestock improvement, Nat. Rev. Genet. 20 (2019) 135–156.
[27] E. Giuffra, C.K. Tuggle, F. Consortium, Functional annotation of animal genomes (FAANG): current achievements and roadmap, Annu. Rev. Anim. Biosci. 7 (2019) 65–88.
[28] A.B. Gjuvsland, J.O. Vik, D.A. Beard, P.J. Hunter, S.W. Omholt, Bridging the genotype-phenotype gap: what does it take? J. Physiol. 591 (2013) 2055–2066.
[29] M.E. Goddard, K.E. Kemper, I.M. MacLeod, A.J. Chamberlain, B.J. Hayes, Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture, Proc. Biol. Sci. 283 (2016).
[30] C. Gross, D. de Ridder, M. Reinders, Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse, BMC Bioinformatics 19 (2018).
[31] C. Gross, M. Derks, H.J. Megens, M. Bosse, M.A.M. Groenen, M. Reinders, D. de Ridder, pCADD: SNV prioritisation in *Sus scrofa*, Genetic Select. Evol. 52 (2020) 4.
[32] D. Habier, R.L. Fernando, D.J. Garrick, Genomic BLUP decoded: a look into the black box of genomic prediction, Genetics 194 (2013) 597.
[33] S.J.G. Hall, Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data, Animal 10 (11) (2016) 1778–1785.
[34] S. Halpain, L. Dehmelt, The MAP1 family of microtubule-associated proteins, Genome Biol. 7 (2006) 224.
[35] J.N. Hellwege, J.M. Keaton, A. Giri, X. Gao, D.R. Velez Edwards, T.L. Edwards, Population stratification in genetic association studies, Curr. Protoc. Human Genet. 95 (2017), 1 22 21-21 22 23.
[36] J. Hong, D. Kim, K. Cho, S. Sa, S. Choi, Y. Kim, J. Park, G.S. Schmidt, M.E. Davis, H. Chung, Effects of genetic variants for the swine FABP3, HMGA1, MC4R, IGF2, and FABP4 genes on fatty acid composition, Meat Sci. 110 (2015) 46–51.
[37] S.E. Hunt, W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard, A. Parton, I.M. Armean, S.J. Trevanion, P. Flicek, et al., Ensembl variation resources, Database J. Biol. Databases Curat 2018 (2018).
[38] M.J. Jang, U.H. Park, J.W. Kim, H. Choi, S.J. Um, E.J. Kim, CACUL1 reciprocally regulates SIRT1 and LSD1 to repress PPARgamma and inhibit adipogenesis, Cell Death Dis. 8 (2017) 3201.
[39] K.S. Kim, N. Larsen, T. Short, G. Plastow, 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 11 (2000) 131–135.
[40] K.S. Kim, J.J. Lee, H.Y. Shin, B.H. Choi, C.K. Lee, J.J. Kim, B.W. Cho, T.H. Kim, Association of melanocortin 4 receptor (MC4R) and high mobility group AT-hook 1 (HMGA1) polymorphisms with pig growth and fat deposition traits, Anim. Genet. 37 (2006) 419–421.
[41] E.F. Knol, B. Nielsen, P.W. Knap, Genomic selection in commercial pig breeding, Animal Front. 6 (2016) 15–22.
[42] T.H. Le, O.F. Christensen, B. Nielsen, G. Sahana, Genome-wide association study for conformation traits in three Danish pig breeds, Genetic Select. Evol. 49 (2017) 12.
[43] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[44] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, Genome Project Data Processing S, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[45] X. Li, S.W. Kim, J.S. Choi, Y.M. Lee, C.K. Lee, B.H. Choi, T.H. Kim, Y.I. Choi, J. J. Kim, K.S. Kim, Investigation of porcine FABP3 and LEPR gene polymorphisms and mRNA expression for variation in intramuscular fat content, Mol. Biol. Rep. 37 (2010) 3931–3939.
[46] Y.H. Liu, W.B. Qi, A. Richardson, H. Van Remmen, Y. Ikeno, A.B. Salmon, Oxidative damage associated with obesity is prevented by overexpression of CuZn- or Mn-superoxide dismutase, Biochem. Biophys. Res. Commun. 438 (2013) 78–83.
[47] J. Ma, J. Yang, L. Zhou, J. Ren, X. Liu, H. Zhang, B. Yang, Z. Zhang, H. Ma, X. Xie, et al., A splice mutation in the PHKG1 gene causes high glycogen content and low meat quality in pig skeletal muscle, PLoS Genet. 10 (2014), e1004710.
[48] M. Mahlapuu, C. Johansson, K. Lindgren, G. Hjalm, B.R. Barnes, A. Krook, J. R. Zierath, L. Andersson, S. Marklund, Expression profiling of the gamma-subunit isoforms of AMP-activated protein kinase suggests a major role for gamma3 in white skeletal muscle, Am. J. Physiol. Endocrinol. Metab. 286 (2004) E194–E200.
[49] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The ensembl variant effect predictor, Genome Biol. 17 (2016) 122.
[50] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics 157 (2001) 1819–1829.
[51] D. Milan, J.T. Jeon, C. Looft, V. Amarger, A. Robic, M. Thelander, C. Rogel-Gaillard, S. Paul, N. Iannuccelli, L. Rask, et al., A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle, Science 288 (2000) 1248–1251.
[52] I. Misztal, A. Legarra, I. Aguilar, Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information, J. Dairy Sci. 92 (2009) 4648–4655.
[53] K.D. Mu, Y. Sun, Y. Zhao, T.X. Zhao, Q. Li, M.L. Zhang, H.T. Li, R. Zhang, C. Hu, C. Wang, 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 47 (2019) 352–364.
[54] L. Nagy, J. Marton, A. Vida, G. Kis, E. Bokor, S. Kun, M. Gonczi, T. Docsa, A. Toth, M. Antal, et al., Glycogen phosphorylase inhibition improves beta cell function, Br. J. Pharmacol. 175 (2018) 301–319.
[55] M. Niwa, Y. Numaguchi, M. Ishii, T. Kuwahata, M. Kondo, R. Shibata, K. Miyata, Y. Oike, T. Murohara, IRAP deficiency attenuates diet-induced obesity in mice through increased energy expenditure, Biochem. Biophys. Res. Commun. 457 (2015) 12–18.

[56] R.D. Palmiter, Reduced levels of neurotransmitter-degrading enzyme PRCP promote obesity, J. Clin. Investig. 119 (2009) 2130–2133.
[57] C. Perleberg, A. Kind, A. Schnieke, Genetically engineered pigs as models for human disease, Dis. Models Mech. (2018) 11.
[58] K. Piorkowska, K. Ropka-Molik, T. Szmatola, K. Zygmunt, 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. 168 (2014) 38–44.
[59] C.P. Ponting, R.C. Hardison, What fraction of the human genome is functional? Genome Res. 21 (2011) 1769–1776.
[60] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.
[61] P. Rentzsch, D. Witten, G.M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome, Nucleic Acids Res. 47 (2019) D886–D894.
[62] H. Reyer, S. Ponsuksili, K. Wimmers, E. Murani, Transcript variants of the porcine glucocorticoid receptor gene (NR3C1), Gen. Comp. Endocrinol. 189 (2013) 127–133.
[63] M. Ron, J.I. Weller, From QTL to QTN identification in livestock - winning by points rather than knock-out: a review, Anim. Genet. 38 (2007) 429–439.
[64] K. Rosenvold, J.S. Petersen, H.N. Laerke, S.K. Jensen, M. Therkildsen, A. H. Karlsson, H.S. Moller, H.J. Andersen, Muscle glycogen stores and meat quality as affected by strategic finishing feeding of slaughter pigs, J. Anim. Sci. 79 (2001) 382–391.
[65] D.J. Schaid, W. Chen, N.B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping, Nat. Rev. Genet. 19 (2018) 491–504.
[66] C.J. Sinal, M. Tohkin, M. Miyata, J.M. Ward, G. Lambert, F.J. Gonzalez, Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis, Cell 102 (2000) 731–744.
[67] J. Sjolund, F.G. Pelorosso, D.A. Quigley, R. DelRosario, A. Balmain, Identification of Hipk2 as an essential regulator of white fat development, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 7373–7378.
[68] M.E. Skinner, A.V. Uzilov, L.D. Stein, C.J. Mungall, I.H. Holmes, JBrowse: a next-generation genome browser, Genome Res. 19 (2009) 1630–1638.
[69] A. Tanabe, T. Yanagiya, A. Iida, S. Saito, A. Sekine, A. Takahashi, T. Nakamura, T. Tsunoda, S. Kamohara, Y. Nakata, 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. 92 (2007) 1145–1154.
[70] J. Ten Napel, M. Calus, M. Lidauer, I. Stranden, E. Mäntysaari, H. Mulder, R. Veerkamp, MiXBLUP, the Mixed-Model Best Linear Unbiased Prediction Software for PCs for Large Genetic Evaluation Systems, 2016. Version.
[71] C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq, Nat. Biotechnol. 31 (2013) 46.
[72] J. Trevaskis, K. Walder, V. Foletta, L. Kerr-Bayles, J. McMillan, A. Cooper, S. Lee, K. Bolton, M. Prior, R. Fahey, 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 146 (2005) 3757–3764.
[73] P. Uimari, A. Sironen, A combination of two variants in PRKAG3 is needed for a positive effect on meat quality in pigs, BMC Genetics 15 (2014).
[74] C. UniProt, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.
[75] M. van Son, N.H. Tremoen, A.H. Gaustad, F.D. Myromslien, D.I. Vage, E. B. Stenseth, T.T. Zeremichael, E. Grindflek, RNA sequencing reveals candidate genes and polymorphisms related to sperm DNA integrity in testis tissue from boars, BMC Vet. Res. 13 (2017) 362.
[76] M. van Son, M.S. Lopes, H.J. Martell, M.F.L. Derks, L.E. Gangsei, J. Kongsro, M. N. Wass, E.H. Grindflek, B. Harlizius, 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. (2019) 10.
[77] R. Veroneze, P.S. Lopes, S.E.F. Guimaraes, F.F. Silva, M.S. Lopes, B. Harlizius, E. F. Knol, Linkage disequilibrium and haplotype block structure in six commercial pig lines, J. Anim. Sci. 91 (2013) 3493–3501.
[78] D. Villar, C. Berthelot, S. Aldridge, T.F. Rayner, M. Lukk, M. Pignatelli, T.J. Park, R. Deaville, J.T. Erichsen, A.J. Jasinska, et al., Enhancer evolution across 20 mammalian species, Cell 160 (2015) 554–566.
[79] A. Warr, N. Affara, B. Aken, H. Beiki, D.M. Bickhart, K. Billis, W. Chow, L. Eory, H. A. Finlayson, P. Flicek, et al., An improved pig reference genome sequence to enable pig genetics and genomics research, Gigascience (2020) 9.
[80] J. Yang, S.H. Lee, M.E. Goddard, P.M. Visscher, GCTA: a tool for genome-wide complex trait analysis, Am. J. Hum. Genet. 88 (2011) 76–82.
[81] S. Yoshizumi, S. Suzuki, M. Hirai, Y. Hinokio, T. Yamada, T. Yamada, U. Tsunoda, H. Aburatani, K. Yamaguchi, T. Miyagi, et al., Increased hepatic expression of ganglioside-specific sialidase, NEU3, improves insulin sensitivity and glucose tolerance in mice, Metabolism 56 (2007) 420–429.
[82] J.Y. Yun, H.G. Jin, Y. Cao, L.C. Zhang, Y.M. Zhao, X. Jin, Y.S. Yu, RNA-Seq analysis reveals a positive role of HTR2A in Adipogenesis in Yan yellow cattle, Int. J. Mol. Sci. 19 (2018).
[83] D.R. Zerbino, P. Achuthan, W. Akanni, M.R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C.G. Giron, et al., Ensembl 2018, Nucleic Acids Res. 46 (2018) D754–D761.
[84] B. Zhang, P. Shang, Y.Z. Qiangba, A.S. Xu, Z.X. Wang, H. Zhang, The association of NR1H3 gene with lipid deposition in the pig, Lipids Health Dis. 15 (2016).
[85] B. Zhang, J. Wu, Y. Cai, M. Luo, B. Wang, Y. Gu, AAED1 modulates proliferation and glycolysis in gastric cancer, Oncol. Rep. 40 (2018) 1156–1164.