Profiling of open chromatin in developing pig (*Sus scrofa*) muscle to identify regulatory regions

Mazdak Salavati[1,2], Shernae A. Woolley[1], Yennifer Cortés Araya[1], Michelle M. Halstead[3], Claire Stenhouse[1,4], Martin Johnsson[1,5], Cheryl J. Ashworth[1], Alan L. Archibald[1], Francesc X. Donadeu[1], Musa A. Hassan[1,2§] and Emily L. Clark[1,2,§]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Edinburgh, UK

[2]Centre for Tropical Livestock Genetics and Health (CTLGH), Roslin Institute, University of Edinburgh, Easter Bush Campus, EH25 9RG, UK

[3]Department of Animal Science, University of California, Davis, California, US

[4]Department of Animal Science, Texas A&M University, Texas, USA

[5]Swedish University of Agricultural Sciences, Uppsala, Sweden

[§]Corresponding authors: emily.clark@roslin.ed.ac.uk ; musa.hassan@roslin.ed.ac.uk

Abstract

There is very little species-specific information about how the genome is regulated in domestic pigs (*Sus scrofa*). This lack of knowledge hinders efforts to define and predict the effects of genetic variants in pig breeding programmes. In order to address this knowledge gap, we need to identify regulatory sequences in the pig genome starting with regions of open chromatin. We have optimised the 'Improved Protocol for the Assay for Transposase-Accessible Chromatin (Omni-ATAC-seq)' to profile regions of open chromatin in flash frozen pig muscle tissue samples. This protocol has allowed us to identify putative regulatory regions in semitendinosus muscle from 24 male piglets. We collected samples from the smallest, average, and largest sized male piglets from each litter through five developmental time points. The ATAC-Seq data were mapped to *Sscrofa11.1* with Bowtie2 and Genrich were used for post-alignment peak-calling. Of the 4661 putative regions of accessible chromatin identified, >50% were within 1 kb of known transcription start sites. The size of each open chromatin region varied according to the developmental time point. At day 90 of gestation, we investigated chromatin openness relative to foetal piglet size. In parallel we measured genome-wide gene expression and allele-specific expression using RNA-Seq analysis of the same muscle samples. We found regions of open chromatin associated with down regulation of genes involved in muscle development in small sized foetal piglets. The dataset that we have generated here provides: i) a resource for studies of genome regulation in pigs, and ii) contributes valuable functional annotation information to filter genetic variants for use in

genomic selection in pig breeding programmes. Future work could leverage the ATAC-Seq data with very large datasets of genetic variants from phenotyped pigs. This approach could inform chromatin aware genomic prediction models and determine whether regions of open chromatin are enriched for trait-linked variants, and especially for muscle and meat traits.

Key words: ATAC-Seq, *Sscrofa11.1*, RNA-Seq, frozen tissue, muscle, allele-specific expression

Introduction

The domestic pig (*Sus scrofa*) is a hugely important farmed animal species globally, contributing a source of healthy animal protein to feed the growing human population. Meeting the increased demand for healthy sustainably produced food from pigs in coming decades will require novel breeding strategies and management practices that will rely on an improved ability to predict phenotype from genotype (Clark et al., 2020). High resolution annotations of the expressed and regulatory regions of farmed animal genomes provides a resource to accurately link genotype to phenotype (Andersson et al., 2015). Variants in putative regulatory regions have been associated with >100 phenotypes in humans (Pai et al., 2015). Recently, a functional regulatory variant in the gene myosin heavy chain 3 (*MYH3*) was shown to influence muscle fibre type composition in Korean native pigs (Cho et al., 2019). There is very little species-specific information, especially, about how the genome is regulated in domestic pigs. This lack of knowledge hinders efforts to identify causative variants for complex traits, and a better knowledge of genome regulation might also improve genomic prediction in breeding programmes. In order to address this knowledge gap, we need to identify regulatory sequences in the pig genome, starting with regions of open chromatin.

Activation of regulatory DNA drives gene expression patterns that influence the phenotypic characteristics of domestic pigs. Measurement of open chromatin gives a quantitative genome wide profile of chromatin accessibility appearing as ‘peaks’ in the data generated for each tissue sample (Thurman et al., 2012). These peaks can reflect the function of the adjoining regulatory DNA (Thurman et al., 2012). The Assay for Transposable Chromatin (ATAC-Seq) (Buenrostro et al., 2015) has been used successfully to profile regions of open chromatin in chicken, cattle and pig genomes (Halstead et al., 2020a, 2020b). In this study we optimised the ‘Improved Protocol for the Assay for Transposase-Accessible Chromatin (Omni-ATAC-seq)’ (Corces et al., 2017) to profile regions of open chromatin in flash frozen pig muscle tissue samples.

Muscle is an important tissue in commercial pig production as muscle traits (e.g. meat and carcass quality) act as economic drivers in pig breeding programmes. Prior to this study knowledge of open chromatin in pig muscle was limited to data from only two adult animals

(Halstead et al., 2020b). For this study, we collected semitendinosus muscle tissues from piglets at five different stages of development (three foetal stages, one neonatal and one juvenile stage). The developmental stages were chosen according to their relevance to hyperplasic muscle development in the foetus and post-natal muscle hypertrophy (Ashmore et al., 1973; Wigmore and Stickland, 1983; Rudar et al., 2019). We hypothesised that gene expression and regulation in semitendinosus muscle tissue would change as the piglets aged, allowing us to identify the transcripts and regions of open chromatin that drive myogenesis. Several studies have profiled gene expression during foetal development in pigs (Zhao et al., 2011, 2015; Yang et al., 2015; Ayuso et al., 2016), but none to date have examined how chromatin openness changes as the piglet develops.

The number of muscle fibres in pigs is proportional to weight at birth (Aiello et al., 2018; Stange et al., 2020). Low birth weight in pigs has been shown to cause lifelong impairments in muscle development and growth (Rehfeldt and Kuhn, 2006). Low birth weight piglets often display ‘catch up’ growth, but at the expense of laying down a higher proportion of body fat compared to normal sized littermates (Estany et al., 2017). Consistent with these observations, mesenchymal stem cells from intrauterine growth-restricted piglets show a differentiation bias towards the adipocyte lineage in comparison with their normal sized litter mates (Weatherall et al., 2020). Low birth weight piglets tend to produce fattier, less valuable carcasses from a production perspective and as such their incidence within pig litters should be kept at a minimum (Pardo et al., 2013). Piglet size variation within a litter is likely to be determined by many different physiological variables including variation in placental blood flow (Stenhouse et al., 2018) but may also be influenced by genomic drivers (Wang et al., 2016; Li et al., 2020).

The study we present here used samples of muscle tissue from a common commercial breed cross (Large White x Landrace) to generate ATAC-Seq and RNA-Seq data from the same individuals to characterise the expressed and regulatory regions of the genome during pig development. The aims of the study were to: 1) Optimise the Omni-ATAC-Seq protocol for frozen pig muscle tissue; 2) Map regions of open chromatin in semitendinosus muscle tissue from small, average and large sized male piglets at five developmental stages (days 45, 60 & 90 pre-natal, one and six weeks post-natal) and 3) Integrate ATAC-Seq peaks with RNA-Seq data to link regions of open chromatin with gene and allele-specific expression patterns. To our knowledge, this is the first time the Omni-ATAC-Seq protocol has been optimised for frozen muscle tissue from a farmed animal species and we have provided a detailed protocol on the Functional Annotation of Animal Genomes (FAANG) data portal (https://data.faang.org). The outcomes of the study will help to: 1) understand the molecular drivers of muscle growth in pigs; 2) provide a foundation for functionally validating target genomic regions *in vitro* and 3) identify high quality causative variants for muscle growth with

the goal of harnessing genetic variation and turning it into sustainable genetic gain in pig breeding programmes.

Methods

Animals

Tissue samples for this study were collected from Large White x Land Race pigs that were euthanized, not specifically for this study but for other on-going projects on the effects of foetal size on pig development at The Roslin Institute. Foetal tissues were collected from pregnant sows that were euthanized with sodium pentobarbitone 20% w/v (Henry Schein Animal Health, Dumfries, UK) at a dose of 0.4 ml/kg by intravenous injection. Post-natal samples were collected after euthanasia by captive bolt. All studies from which samples were collected were approved by the University of Edinburgh and Roslin Institute Animal Ethics and Welfare Review Board (AWERB) and were in accordance with the U.K. Animals (Scientific Procedures) Act, 1986.

Sample collection of frozen muscle tissue samples for ATAC-Seq and RNA-Seq

The tissue samples used for this study were from archived material (with the exception of the piglets that were six weeks of age) collected from the largest, smallest and average sized male piglets per litter at five different developmental stages (Table 1). The largest, smallest and average sized piglets per litter were selected according to body weight for the foetal time points and birth weight for post-natal time points (Supplementary Table 1). Developmental stages were chosen, according to previous studies (Ashmore et al., 1973; Wigmore and Stickland, 1983; Rudar et al., 2019) as follows:

1. Day 45 of gestation - when primary muscle fibres form.
2. Day 60 of gestation - when secondary muscle fibres begin to form.
3. Day 90 of gestation - when fibre formation ceases after which subsequent muscle growth occurs through fibre hypertrophy.
4. One week of age - during active muscle hypertrophy.
5. Six weeks of age – once muscle hypertrophy has levelled off.

Due to limited sample availability the experimental design is unbalanced (Table 1), specifically for days 45 (N=3) and 60 (N=3) of gestation, and six weeks of age (N=5), only one complete set of littermates (smallest, average and largest) was included in the analysis. At day 90 (N=11) of gestation three complete litters and one incomplete litter were included, while at one week of age samples were only available from the smallest and largest piglets from

one litter (N=2). We have included a flow chart describing which samples were analysed at each stage of this study (Figure 1).

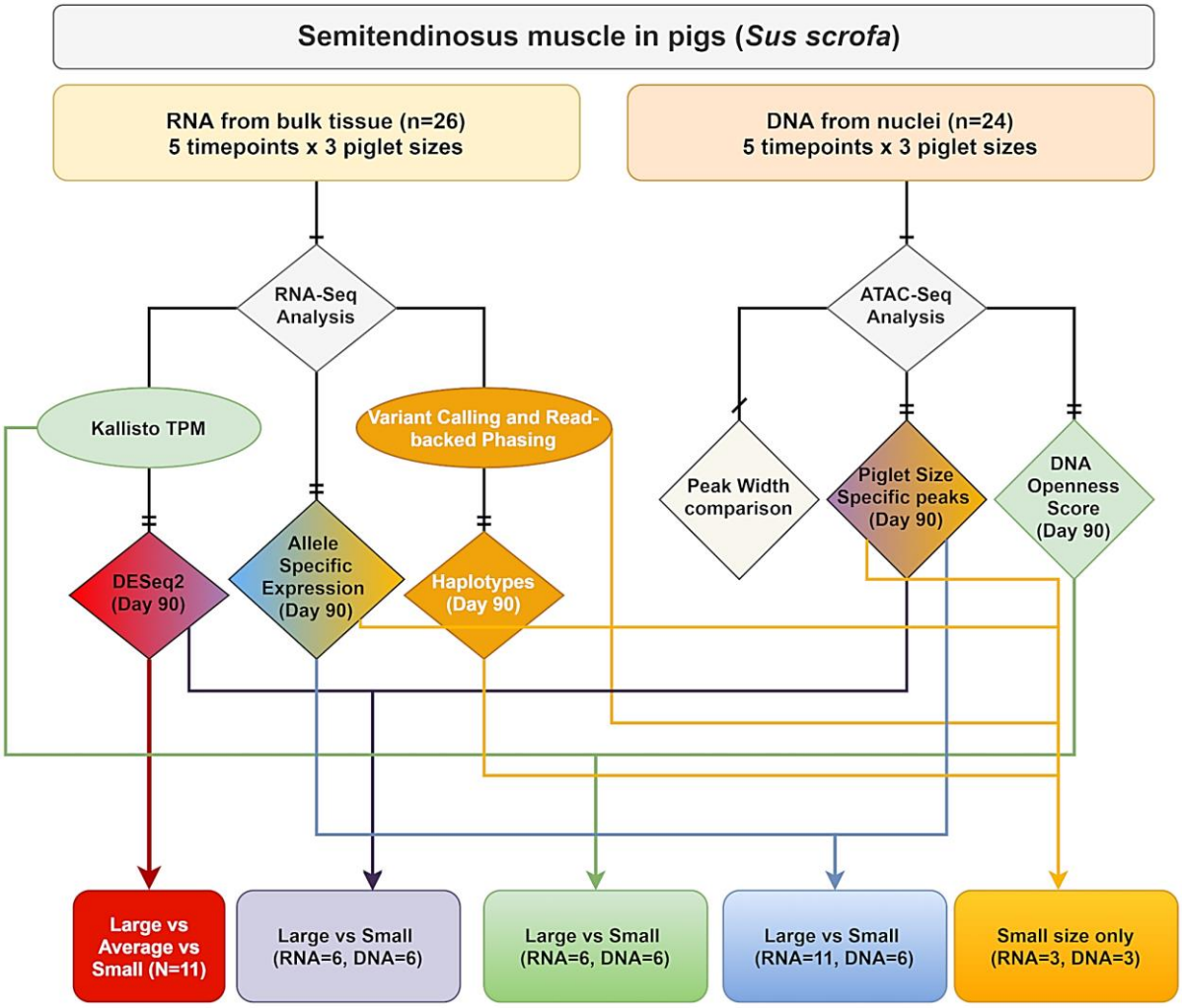


Figure 1: Flowchart describing the experimental design and samples included in each stage of the analysis performed in this study. Colour coding indicates where there are overlaps in the analysis performed for each component of the study.

Samples were collected from the semitendinosus muscle from the hind leg of piglets from each developmental stage (Table 1). The only exception was day 45, when whole muscle tissue was collected, because it was not possible to differentiate specific muscle types at this early stage of development. Each sample was flash frozen in liquid nitrogen, as quickly as possible within an hour post euthanasia and stored at -80°C for future analysis. From the six week old piglets additional samples were collected in sucrose buffer to isolate and cryopreserve nuclei according to the method described in (Halstead et al., 2020a). Cryopreserving isolated nuclei for a small number of samples would allow us to validate the data we generated from the flash frozen material, which we were optimising for the first time

on muscle tissue. The protocol for collection of tissue samples at the farm is available via the FAANG Data Coordination Centre https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_Collection_of_tissue_samples_for_ATAC-Seq_and_RNA-Seq_from_large_animals_20200618.pdf.

Table 1. Details of muscle tissues sampled, for ATAC-Seq and RNA-Seq, from piglets at five developmental stages.

Sample ID	Litter Identifier	Gestational Age in Days	Tissue Sampled	Piglet Size	ATAC-Seq	RNA-Seq
Pre-natal time points						
D45L220716	220716	45	Hind Leg Muscle	Largest	Yes	Yes
D45N220716	220716	45	Hind Leg Muscle	Average	Yes	Yes
D45S220716	220716	45	Hind Leg Muscle	Smallest	Yes	Yes
D60L120916	120916	60	Semitendinosus Muscle	Largest	Yes	Yes
D60N120916	120916	60	Semitendinosus Muscle	Average	Yes	Yes
D60S120916	120916	60	Semitendinosus Muscle	Smallest	Yes	No
D60S23976	23976	60	Semitendinosus Muscle	Smallest	No	Yes
D90L251016	251016	90	Semitendinosus Muscle	Largest	Yes	Yes
D90L111016	111016	90	Semitendinosus Muscle	Largest	Yes	Yes
D90L121016	121016	90	Semitendinosus Muscle	Largest	Yes	Yes
D90N251016	251016	90	Semitendinosus Muscle	Average	Yes	Yes
D90N111016	111016	90	Semitendinosus Muscle	Average	Yes	Yes
D90N031115	031115	90	Semitendinosus Muscle	Average	Yes	Yes
D90N121016	121016	90	Semitendinosus Muscle	Average	Yes	Yes
D90N231115	231115	90	Semitendinosus Muscle	Average	Yes	Yes
D90S231115	231115	90	Semitendinosus Muscle	Smallest	Yes	Yes
D90S251016	251016	90	Semitendinosus Muscle	Smallest	Yes	Yes
D90S111016	111016	90	Semitendinosus Muscle	Smallest	Yes	Yes
D90S121016	121016	90	Semitendinosus Muscle	Smallest	Yes	Yes
Post-natal time points		**Age in weeks**				
1WKL100918	100918	One	Semitendinosus Muscle	Largest	Yes	Yes
1WKS100918	100918	One	Semitendinosus Muscle	Smallest	Yes	Yes
6WKA050219	050219	Six	Semitendinosus Muscle	Average	Yes	Yes
6WKS050219	050219	Six	Semitendinosus Muscle	Smallest	Yes	Yes
6WKS131218	131218	Six	Semitendinosus Muscle	Smallest	No	Yes
6WKA131218	131218	Six	Semitendinosus Muscle	Average	No	Yes
6WKL131218	131218	Six	Semitendinosus Muscle	Largest	No	Yes
Cryopreserved Nuclei Preparations		**Age in weeks**				
6WKA050219CN	050219	Six weeks	Semitendinosus Muscle	Average	Yes	No
6WKS050219CN	050219	Six weeks	Semitendinosus Muscle	Smallest	Yes	No

Isolation of cryopreserved nuclei from fresh muscle tissue and preparation of tagmented nuclear DNA

We used the protocol described in (Halstead et al., 2020a) to isolate and cryopreserve intact nuclei from fresh muscle tissue samples from the six week old piglets (Table 1). Briefly, each

tissue sample was transferred to a GentleMACS C tube (Mitenyi Biotec, Germany) with sucrose buffer and dissociated. The homogenate was then filtered and Dimethyl Sulfoxide (DMSO) (Sigma Aldrich, USA) added (10% final concentration), before freezing at -80°C overnight in a Mr Frosty (Nalgene, USA), then transferring to a -80°C freezer for long-term storage. The full protocol for preparation of cryopreserved nuclei from fresh muscle tissue is available via the FAANG Data Coordination Centre https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_Cryopreservation_of_Nuclei_for_ATACSeq_using_GentleMACS_20201119.pdf.

To prepare tagmented DNA the cryopreserved nuclei preparations were thawed slowly at room temperature by adding 500µl of cold 1x Phosphate Buffered Saline (PBS), filtered then centrifuged at 500 x g at 4ºC in a swinging bucket centrifuge for 10 minutes. After centrifugation, the pellet was resuspended in 1 ml cold ATAC-seq RSB buffer + 0.1%Tween20 (Sigma Aldrich, USA) for lysis and centrifuged for 10 minutes at 500g at 4ºC. The pellet of nuclei was then washed in PBS and resuspended in 50 µl transposition mix (25 µl TD buffer, 2.5 µl TDE1 enzyme, Molecular Biology Grade Sterile H_2O) from the Nextera DNA Sample Prep Kit (Ilumina, USA). The pellet was incubated with the transposition mix for 60 minutes at 37ºC at 300 rpm on a thermomixer. The pellet of transposed nuclear DNA, was purified with a MinElute PCR purification kit (Qiagen, Germany), eluted in 15µl of Buffer EB, and stored at -20ºC. The full protocol is available via the FAANG Data Coordination Centre https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_ATAC-Seq_DNAIsolationandTagmentation_Cryopreserved_Muscle_Nuclei_Preparations_20200720.pdf

Isolation of nuclei from frozen muscle tissue and preparation of tagmented nuclear DNA

ATAC-Seq libraries were prepared using a version of the 'Improved Protocol for the Assay for Transposase-Accessible Chromatin (Omni-ATAC-seq)' (Corces et al., 2017) which we optimised for flash frozen pig muscle tissue samples for this study. The main modification that we introduced to the protocol was an initial dissociation step using a GentleMACS Dissociator (Mitenyi Biotec, Germany), essentially combining the Omni-ATAC-seq protocol with the initial steps from (Halstead et al., 2020a). The protocol is described in full at https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_ATAC_Seq_DNAIsolationandTagmentation_Frozen_Muscle_Tissue_20200720.pdf and summarised here. The components of each of the buffers are included in Supplementary Table 2. Each flash frozen tissue sample (~200mg per sample) was chopped into small pieces over dry ice and then dissociated in a GentleMACS C-tube (Mitenyi Biotec, Germany) in 1ml of 1XHB buffer (+Protease Inhibitor Cocktail (PIC)). The samples were dissociated using programme m_muscle_0.1_0.1 (equivalent to 'E0.1c Tube') twice on a GentleMACS Dissociator (Mitenyi Biotec, Germany).

Immediately after dissociation the samples were filtered through a 70µm corning cell strainer (Sigma Aldrich, USA) then centrifuged at 3000 x g for 5 minutes. The pellet was resuspended in 400µl 1XHB buffer and transferred to a 2ml Eppendorf Protein Lo-Bind tube (Eppendorf, UK). 400µl of 50% Iodixanol solution (Opti-Prep Density Gradient Medium) (SLS, UK) was added to the 400µl of cell solution (final 25% Iodixanol). An Iodixanol gradient was then created and samples transferred to a swinging bucket centrifuge and spun for 25 minutes at maximum speed at 4ºC with no brake. A thin "whitish" band appeared between layers two and three of the gradient. Evaluation and counting of nuclei was performed by staining with Trypan Blue (ThermoFisher Scientific, USA). 1ml of ATAC-RSB Buffer + 0.1%Tween20 (Sigma Aldrich, USA) was then added to lyse the nuclei and the sample centrifuged for 10 minutes at 500 x g at 4°C. The pellet was then gently resuspended in 50 µl transposition mix for tagmentation as described for cryopreserved nuclei samples above.

ATAC-Seq library preparation

The library preparation protocol, adapted from (Corces et al., 2017)**,** was used for the flash frozen tissues and the cryopreserved nuclei preparations. The protocol described in full is available via the FAANG Data Coordination Centre https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_ATAC-Seq_LibraryPreparationandSizeSelection_20200720.pdf. A PCR reaction mix was set up comprising 10ul molecular biology grade H_2O, 2.5µl Ad1 primer 25µM, 2.5µl Ad2.x primer 25µM (variable index see Supplementary Table 3), and 25µl 2x NEBNext Hi-Fi PCR mix (NEB, USA) per reaction. 10µl of transposed DNA was added to each reaction and 5 amplification cycles of the following PCR reaction performed: 72°C for 5 min, 98°C for 30 sec, 98°C for 10 sec, 63°C for 30 sec, 72°C for 1min. The GreenLeaf Quantitative PCR (qPCR) Protocol (Buenrostro et al., 2015) was used to determine the number of additional PCR amplification cycles that were required for each sample, to stop amplification prior to saturation and avoid variation across samples caused by PCR bias. Samples for which more than 5-7 additional cycles were required were discarded due to the high probability of PCR bias caused by additional cycles. Amplified ATAC-Seq libraries were then purified with a MinElute PCR purification kit (Qiagen, Germany). Library quality was checked on the Agilent 2200 TapeStation System (Agilent Genomics, USA). Libraries were assessed for quality according to an even distribution of fragments and a clearly differentiated sub-nucleosomal fragment as described in (Halstead et al., 2020a)**.** If library quality was sufficient the sub-nucleosomal fragment (150-250bp) was size selected, in order to minimise the signal to noise ratio, as suggested in (Halstead et al., 2020a). Size selection was performed using a Thermo Scientific E-Gel System (ThermoFisher Scientific, USA). To check the size of the selected fragment an

aliquot was run on the Agilent 2200 TapeStation System (Agilent Genomics, USA). After size selection the libraries were pooled and stored at -20°C prior to sequencing.

Sequencing of ATAC-Seq libraries

Pooled libraries (4 batches) were sequenced to generate 50nt paired-end reads on an Illumina NovaSeq 6000 platform using a single S2 flow cell. The majority of libraries generated >90M paired-end reads (Min: 9.8e+07, Max: 3.5e+08, Median: 1.97e+08).

ATAC-Seq data processing and mapping

Quality control of the raw sequence data was performed using FastQC v0.11.9 and multiQC v1.9 (Ewels et al., 2016). Quality control metrics included the rate of PCR duplication. The paired end reads were trimmed using Trimmomatic v0.39 (Bolger et al., 2014). The trimmed reads were then mapped to the *Sscrofa11.1* pig reference genome (Warr et al., 2020) available from Ensembl (GCA_000003025.6) using Bowtie2 v2.3.5.1 and the default flags of the *--very-sensitive* mode. The BAM files that were generated were then sorted and indexed using samtools v1.6 (Li, 2011). Overall on average more than 75M reads per samples were uniquely mapped (Min: 2.47e+07 ,Max: 1.28e+08, Median: 7.74e+07, Mean ± SD: 7.72e+07 ± 3.15e+07). The PCR duplication level was 43% ± 8 (mean ± SD) on average across all libraries.

ATAC-Seq peak calling using Genrich

ATAC-Seq peak calling for each timepoint was performed using Genrich v0.5 (John M. Gaspar, 2020) under ATAC-Seq mode. A set of peaks that were shared across the five developmental timepoints was produced using bedtools v2.26.0 (Quinlan and Hall, 2010) with the intersect functionality series of scripts. All the scripts used for this analysis can be found in Supplementary File 1 and the code repository https://msalavat@bitbucket.org/msalavat/pig_muscle.git. Briefly a merger of all peaks that were present in at least one time point was performed to create a set of peaks that were shared across all time points. Then, peaks that were specific to each time point were subtracted from the merger. The BED files of peaks specific to each developmental time point are available in Supplementary File 1.

For the Day 90 samples only, we re-analysed the ATAC-Seq peaks in the three sizes of foetal piglet per litter (large, average and small). The peak calling was performed using the same Genrich flags as previously described, and we separated peaks shared between all size classes and size class specific peaks with bedtools. The size specific peaks generated for the foetal piglets at Day 90 of gestation and the scripts used to produce them can also be found

in Supplementary File 1 and the code repository https://msalavat@bitbucket.org/msalavat/pig_muscle.git.

Multidimensional scaling analysis for comparison of ATAC-Seq libraries

Non-linear multidimensional scaling (NMDS) of the ATAC-Seq libraries was performed using the MASS::IsoMDS package (Venables and Ripley, 2002) to ensure that there were no obvious outlying samples and that tissues of the same type clustered together in a biologically meaningful manner. A distance matrix (Manhattan distance) was produced using a TSS annotated (±3kbp) set of ATAC-Seq peaks from each of the samples. The distance matrix was then processed for multidimensional scaling. NMDS was also used to compare the ATAC-Seq libraries prepared from either flash frozen muscle tissue or cryopreserved nuclei from two 6-week-old piglets.

RNA isolation and quality control

The RNA isolation protocol is described in full at https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_RNA_IsolationoftotalRNAfromfrozentissuesamples_20200720.pdf . RNA was extracted from approximately 60mg of tissue. Tissue samples were homogenised in 1ml of TRIzol (Thermo Fisher Scientific, USA) with CK14 (VWR, USA) tissue homogenising ceramic beads on a Precellys Tissue Homogeniser (Bertin Instruments; France) at 5000 rpm for 20 sec. RNA was then isolated using the TRIzol protocol (Thermo Fisher Scientific, USA) and column purified to remove DNA and trace phenol using a RNeasy Mini Kit (Qiagen, Germany) following the manufacturer's instructions. RNA integrity (RIN^e) was estimated on an Agilent 2200 TapeStation System (Agilent, USA) to ensure RNA quality was of $RIN^e > 7$. RIN^e and other quality control metrics for the RNA samples are included in Supplementary Table 4.

Poly-A enriched library preparation and sequencing

Strand-specific paired-end reads with a fragment length of 100bp for each sample were generated by Edinburgh Genomics, using the Illumina TruSeq mRNA library preparation protocol (poly-A selected) (Illumina; Part: 15031047 Revision E). mRNA-Seq libraries were sequenced on an Illumina NovaSeq 6000 platform to generate >90M paired end reads per sample (Min: 6.6e+07, Max:1.21e+08, Mean: 9.17+e07).

RNA-Seq data analysis workflow

The raw sequence data were quality controlled and trimmed using Trimmomatic (Bolger et al., 2014). The kallisto aligner (Bray et al., 2016) was used for expression quantification of the RNA-Seq data. Briefly, a reference transcriptome fasta file of coding sequences was obtained

from *Sscrofa11.1* Ensembl v100 to build a kallisto index file using default settings. The trimmed reads were then mapped for transcript level expression quantification (*de novo)* in kallisto with *--bias* option activated. The output tab separated value files were then imported to R using txImport package (Soneson et al., 2016) for further analysis and visualisations.

The TPM expression estimates for each sample were investigated using principal component analysis (PCA) in FactoMineR to identify any spurious samples that did not cluster as expected (Lê et al., 2008). Differential expression (DE) analysis was performed only on the three sizes of foetal piglet (small, average and large) at Day 90 of gestation. The Likelihood Ratio Test (LRT) model of DESeq2, including post-hoc analysis, was used with small size as the reference level i.e. Denominator in log2FC (*DESeqDataSetFromTximport(txi = dds,design = ~ Size)*). After multiple correction of p values using the BH method (Benjamini and Hochberg, 1995), a false discovery rate of 10% was considered as the significance threshold.

Allele-specific expression analysis workflow

Prior to allele-specific expression (ASE) analysis of the RNA-Seq data, a reference mapping bias correction was performed as previously described by (Salavati et al., 2019). Briefly a *Sscrofa11.1* genome fasta file, GTF gene track and dbSNP VCF file were obtained from Ensembl v100. The trimmed reads were initially mapped to *Sscrofa11.1* using HISAT2 v2.1.0 (Kim et al., 2015). The mapped BAM files were then put through the mapping reference bias removal steps of WASP v0.3.4 (van de Geijn et al., 2015). After the removal of biased mapped reads, the resulting BAM files were processed in GATK v.4.10.1 (McKenna et al., 2010; Van der Auwera et al., 2013) with the ASEReadCounter tool to provide allelic read counts. The allelic counts were then processed in R using GeneiASE (Edsgärd et al., 2016) static mode to screen for genes with significant allelic imbalance given all the heterozygous sites within their coordinates. The mean allelic imbalance (Liptak score produced by GeneiASE output) across biological replicates for each gene was then visualised in R (R Core Team, 2017). In order to be considered robust, an ASE positive gene needed to be present in at least 75% of all biological replicates (for a group size of n=4). Heterozygote variants residing within genomic coordinates of ASE positive genes were labelled as ASE variants for downstream analysis.

Overlay of differentially expressed genes and ATAC-Seq peaks

An overlay of genes that were differentially expressed between the large and small sized foetal piglets at Day 90 was performed using ATAC-Seq peaks within 10Kbp vicinity of the differentially expressed genes (either upstream or downstream). This overlay would show us which of the differentially expressed genes had an ATAC-Seq peak in their vicinity and whether that peak was present in both large and small sized foetal piglets, or only in one of the two sizes. The distance from the start of the gene model to the start of the ATAC-Seq peak

was used as a coordinate system (i.e. positive values meant the peak was either within the gene or within the 3' 10kb upstream region of the gene, and negative values corresponded to 10kb from the 5' end of the gene).

Overlay of allelic imbalance and ATAC-Seq profiles

An overlay of genes exhibiting significant allele specific expression (ASE) with the ATAC-Seq peaks present only in either small or large sized foetal piglets was also performed. This overlay would show us whether presence of an ATAC-Seq peak was correlated with ASE in each of the two sizes of foetal piglet. Allelic imbalance exhibited in a gene needed to be present in ≥75% of biological replicates for it to be considered significant. This criterion was defined as follows:

$$representation\ \% = \frac{Animals\ expressing\ ASE\ gene}{All\ the\ animals\ in\ the\ phenotype\ group} \times 100$$

The ATAC-Seq peaks specific to the large or small sized foetal piglets associated with genes exhibiting significant ASE were overlaid using custom track visualisation in R with ggplot2.

Linking transcript expression level and DNA openness

To investigate whether the expression level of a transcript could be linked to overlap with an ATAC-Seq peak we used calculated a DNA openness score (DOS). For each transcript the exonic regions, 2Kbp upstream of the first exon and 200bp downstream of the coordinates of the last exon were compiled as a 'Genomic Ranges' object in R [GenomicFeatures v.3.11, GenomicRanges package v.3.11 (Lawrence et al., 2013)]. This object (here after referred to as exon_promoter) was overlaid with the ATAC-Seq peak genomic coordinates using the mergeByOverlaps function of the IRanges package v.3.11 (Lawrence et al., 2013) and the following flags (*subject = ATAC, query = exon_promoter, maxgap = 1e4, type = "any"*). The resulting query was used to calculate the DOS, which measures the fraction of the total width of ATAC-seq peaks overlapping a transcript (exonic regions) and its promoter to the total length of the exonic and promoter regions, for each transcript:

$$DOS_{i..j} = \frac{sum\left(ATAC\ peak\ widths\ tx_{i..j}\right)}{length\left(exon_promoter\ tx_{i..j}\right)}$$

The DOS range was then standardized in each sample using the following equation:

$$DOS = \frac{DOS_{i..j} - \min\left(DOS_{i..j}\right)}{max\left(DOS_{i..j}\right) - \min\left(DOS_{i..j}\right)}$$

$$0 < DOS \leq 1$$

RNA-Seq variant calling and read backed haplotype phasing

The RNA-Seq dataset was mapped against *Sscrofa11.1* for the ASE analysis pipeline described above. The BAM files from the Day 90 small sized foetal piglets were used for variant calling. A GATK variant calling workflow was used in order to discover transcriptomic variants in each of the three biological replicates. Briefly BAM files with marked duplicated reads were passed through Picardtools v2.23.0 for addition of read group information (RGID, RGSM, RGLB, RGPL and RGPU SAM flags (Li et al., 2009)). The HaplotypeCaller tool of GATK v.4.10.1 (McKenna et al., 2010; Van der Auwera et al., 2013) was used to produce genomic block resolution GVCF files (*-ERC GVCF*). This produced three GVCF files, one for each foetal piglet, that were then combined using the CombineGVCFs tool in GATK v4.10.1 and passed through GenotypeGVCFs tool to call one set of genotypes for the three small size foetal piglet muscle samples at Day 90. After filtering out the INDEL variants by using SelectVariants tool (-select-type SNP) a hard filtration was performed as described in Supplementary file 2.

The variants that passed all filtration criteria were taken forward to a haplotype phasing step using a WhatsHap v.1.0 (Patterson et al., 2015; Martin et al., 2016) pipeline. The readback phasing of haplotype blocks was performed using the default parameters of the *whatshap phase* command. The resulting vcf file was converted to diploid transcriptome fasta files using BCFtools v.1.9 (Narasimhan et al.). These phased sets of haplotype fasta files (one set for each of the small sized foetal piglets) were used for ORF (open reading frame) prediction analysis of potential transcript isoforms in the 17 genes that showed significant allelic imbalance from the ASE analysis described above. The ORF prediction was performed using the predORF tool of systemPipeR v1.6.2 (Tyler and Girke, 2016) in R. The fasta files and predicted ORFs are included in Supplementary File 2_and Supplementary File 3, respectively. A GTF format track of the phased haplotype blocks was also produced using the *whatshap stats* tool for visualisation purposes (Supplementary File 2). Variants exhibiting ASE that corresponded to the phased haplotype blocks and the ATAC peak peaks for each of the Day 90 small sized foetal piglets were then visualised in IGV v.2.8.9 (Thorvaldsdóttir et al., 2013).

Statistical analysis software and packages

All data analysis for this study was performed via bash scripting and use of R (R Core Team, 2017) on the University of Edinburgh research computing facility (Edinburgh, 2020). The data analysis protocol for ATAC-Seq and RNA-Seq are available at https://data.faang.org/api/fire_api/analyses/ROSLIN_SOP_ATAC-Seq_analysis_pipeline_20201113.pdf and

https://data.faang.org/api/fire_api/analyses/ROSLIN_SOP_RNA-Seq_analysis_pipeline_20201113.pdf .

Results

ATAC-Seq data from frozen pig muscle tissues

ATAC-Seq libraries from four different batches (24 samples in total) were multiplexed and sequenced to achieve 2.02e+08 average reads per sample (Min: 9.8e+07, Max: 3.5e+08, Median: 1.97e+08). Reads were evenly distributed between barcodes across the first three batches. The fourth batch, which included only two samples that had a higher concentration of starting DNA, resulted in more reads per sequencing run compared to the other 22 samples (details in Supplementary File 4). Average chromosomal coverage across autosomes was 7.2x, 4x for X ,3.2x for Y and 4.49e+04 for the mitochondrial chromosome (Supplementary Figure S1).

Multidimensional scaling analysis of the ATAC-Seq libraries from frozen tissue samples

Non-linear multidimensional scaling (NMDS) was used to ensure that the ATAC-Seq dataset was biologically meaningful, reproducible and there were no outlying samples (i.e. samples from the same developmental stage should have a similar peak distribution and cluster together) (Supplementary Figure S2). The ATAC-Seq peak regions in each sample were annotated using the nearest transcript ID and the area under curve of the peak (AUC calculated by Genrich algorithm) was used as the input matrix for the NMDS. The sample from the smallest foetal piglet at Day 45 and 1 week old were separate from the samples from the average and largest piglets of the same age based on the first two components of the NMDS (Supplementary Figure S2). Samples from six-week-old piglets were also relatively separated, while samples from Day 60 and Day 90 foetal piglets clustered together (Supplementary Figure S2).

Multidimensional scaling analysis of ATAC-Seq libraries prepared from either flash frozen muscle tissue or cryopreserved nuclei

Non-linear multidimensional scaling (NMDS) was also used to compare ATAC-Seq libraries prepared from either flash frozen muscle tissue or cryopreserved nuclei from two 6-week-old piglets. The two libraries prepared for the cryo-preserved nuclei samples clustered closely with the libraries prepared for flash frozen tissue indicating there was little difference in the data generated by the two protocols (Supplementary Figure 2). Other metrics, including the percentage of ATAC-Seq peaks within genes, exons and mitochondria were also used to

compare the libraries prepared from cryopreserved nuclei and flash frozen tissue (Supplementary Figure S3). For each of the metrics chosen libraries prepared from cryopreserved nuclei and flash frozen tissue appeared broadly similar, with a slightly higher percentage of ATAC-Seq peaks in exons and in mitochondria in the libraries prepared from cryopreserved nuclei (Supplementary Figure S3).

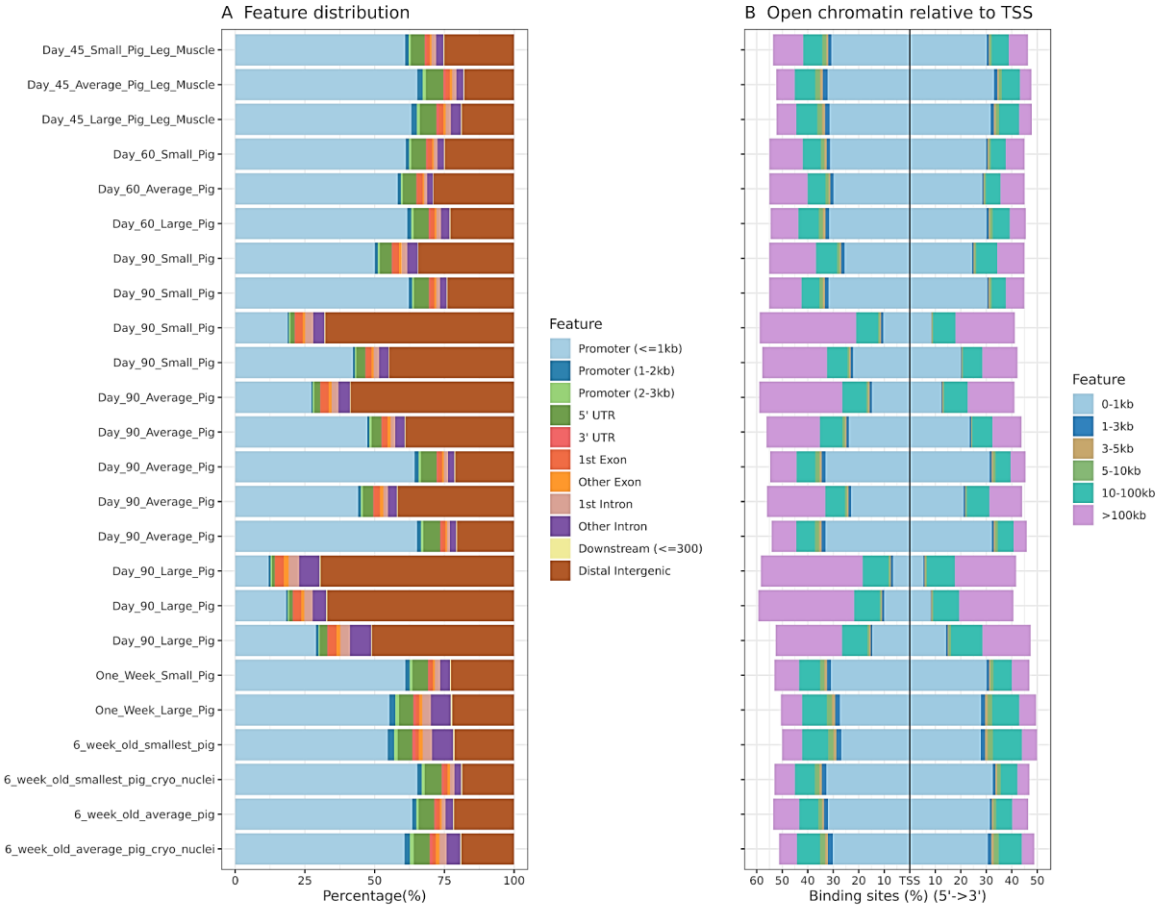


Figure 2: Location of ATAC-Seq peaks within genomic features (A) and distance of ATAC-Seq peaks from the TSS (B) in all 24 samples included in the study. The samples are sorted by the developmental timeline (Day 45 to 6 weeks old from top to bottom) in both panels. The cryopreserved nuclei samples are only present in the 6 week old group and are labelled as cryo_nuclei.

Distribution of ATAC-Seq peaks within genomic features

The feature distribution of the ATAC-Seq peaks in all 24 samples is shown in Figure 2. On average > 6500 peaks (including overlapping regions) were called in each sample (Min: 2.34e+03, Max: 1.42e+04,Median: 6.38e+03, Mean ± SD: 6.72e+03 ± 3.63e+03). More than 52% of the peaks were located in promoter regions in the majority of samples (19/24). There was a slight negative trend between increase in library size (depth of sequencing) and decrease in the number of the peaks called (linear regression: slope= -6e-05 and R^2 = 0.22).

Detailed metrics can be found in Supplementary File 4. In five samples (Day 90 Large [n=3], Day 90 Average and a Day 90 Small) the majority of peaks were located in distal intergenic regions (Figure 2A). We could not find any batch effect, in nuclei extraction or library preparation that might account for this and as such concluded that this variation was related to the samples themselves. There was also little observable difference in how the ATAC-Seq peaks were distributed within the genome of the libraries from cryopreserved nuclei relative to the libraries from flash frozen tissue (Figure 2A-B). The breakdown of genomic feature categories in which peaks were located is presented in Table 2.

Table 2. The frequency of ATAC-Seq peaks in each genomic feature category annotated by ChipSeeker and averaged across samples (total of 3766 annotated peaks from 4661 genomic regions).

Features	***Frequency (% mean ± SD)***
Promoter (1-2Kbp)	1.08 ± 0.52
Promoter (<=1Kbp)	**50.05 ± 16.96**
Promoter (2-3Kbp)	0.74 ± 0.28
5' UTR	4.27 ± 1.64
3' UTR	0.45 ± 0.25
1st Exon	2.07 ± 0.59
Other Exon	1.08 ± 0.32
1st Intron	3.77 ± 0.78
Downstream (<=3Kbp)	0.23 ± 0.08
Distal Intergenic	**34.33 ± 17.29**

Proximity of ATAC-Seq peaks to transcription starts sites (TSSs)

The majority of mapped ATAC-Seq reads aligned within ±1kbp of the transcription start sites (TSSs) as shown in Supplementary Figure S4. The relative distance of the ATAC-Seq peaks from annotated TSSs in the Ensembl *Sscrofa11.1* gene models is shown in Figure 2B. As noted above for genomic feature distribution there was no observable difference in how the ATAC-Seq peaks were positioned relative to the TSS in the libraries from cryopreserved nuclei relative to the libraries from flash frozen tissue (Figure 2B).

Changes in ATAC-Seq peak width distribution during development

Peak width analysis was performed using Genrich for each developmental time point in the following groups: gestational day 45 (Day45 n=3); gestational day 60 (Day60 n=3); gestational D90 (Day90 n=12); neonatal 1 week old (1wo n=2) and 6 weeks old (6wo n=4) (Figure 3). A

density distribution of the width of the ATAC-Seq peaks and the median width size for each timepoint is included in Figure 3A. The Day 90 samples showed significantly smaller peak width in comparison to the other developmental timepoints. In addition to reduced DNA openness (median 194bp vs >450bp in other timepoints), a double peak was present in the Day 90 samples (Figure 3A, 3rd row). This bimodal distribution indicated there was an increased incidence of narrow (~194bp) peaks at the 90 day time point, in addition to the larger size (>450 bp) peaks observed at the other time points. Intersection analysis was performed to identify a set of ATAC-Seq peaks that were specific to each developmental time point (i.e. peaks that were specific to each time point (e.g. Day 90) and shared by none of the other time points) (Figure 3B). The ATAC-Seq peak calls in BED format are included in Supplementary File 1.

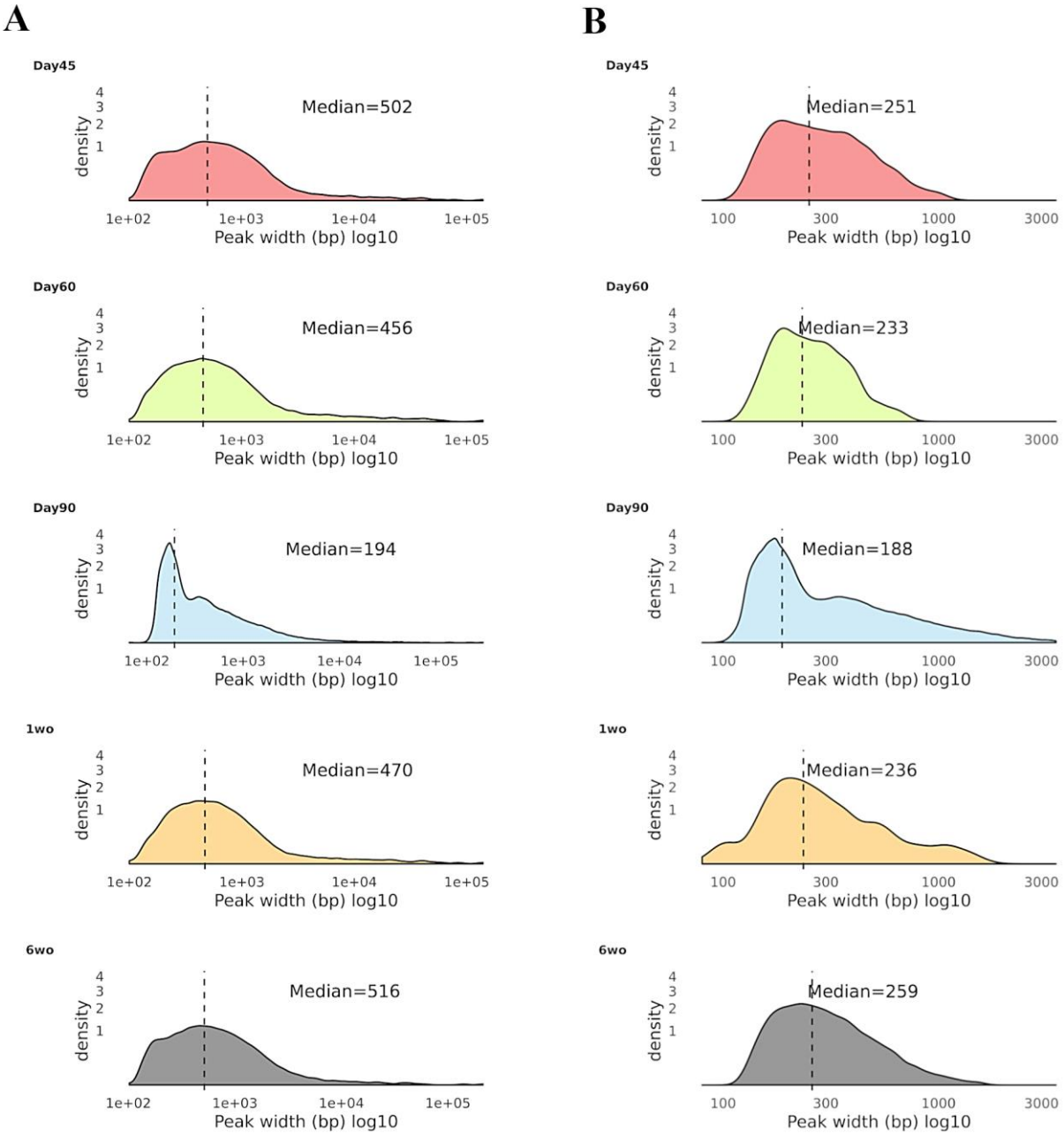


Figure 3: The width distribution of ATAC-Seq peaks at each developmental time point. The dotted line represents the median width for each time point. A) Width of ATAC-Seq peaks shared across all time points, B) Width of ATAC-Seq peaks that were specific to each time point. WO stands for week old.

Read count frequency around the TSSs also varied between the five developmental time points (Figure 4). Read count frequency around the TSSs increased between Day 45 and Day 60 but was reduced by Day 90 (Figure 4). This was followed by a gradual increase in the read count frequency around TSSs in the samples from the one and six-week-old piglets. At Day 90 the read count frequency was greatest in intergenic regions.

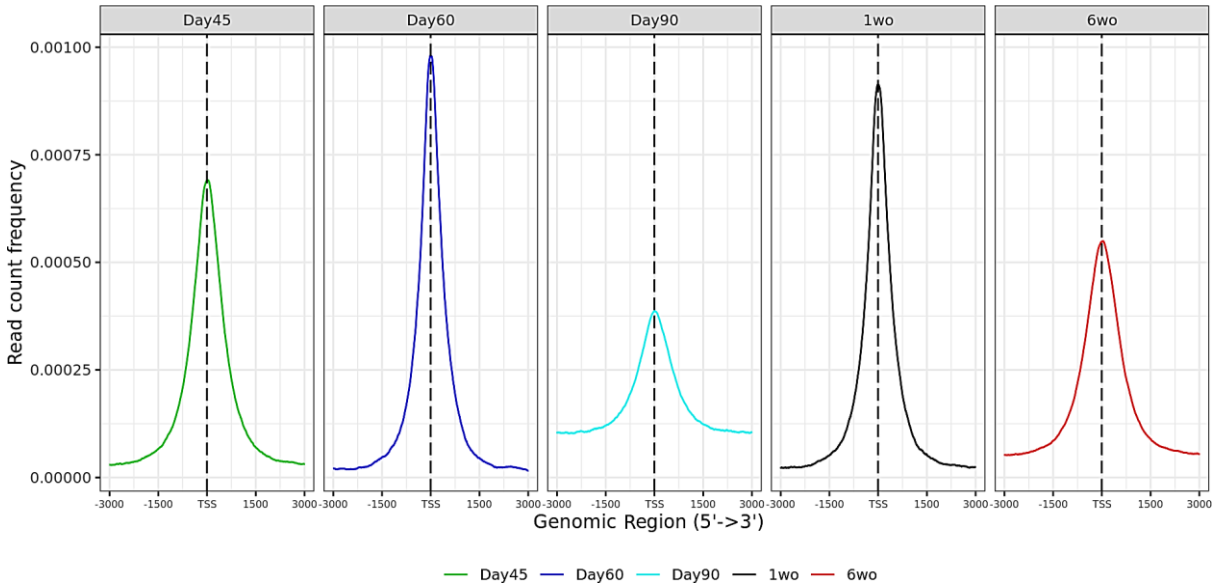


Figure 4: Mapping of ATAC-Seq peaks relative to the TSS, averaged across the three sizes of piglet, showing variation in peak width across the five developmental time points analysed.

Piglet size specific differences in ATAC-Seq peaks at day 90 of gestation

A further comparison of ATAC-Seq peaks at Day 90 of gestation from foetal piglets of large (n=3), average (n=5) and small (n=4) body size was performed. A size specific set of peaks, present only in each size of foetal piglet and not shared across sizes (e.g. Day 90 Large_specific) were identified (Figure 5). The data for size-specific peaks for small and large piglets only represented a small percentage of the total peaks and the data is quite noisy (Figure 5). From a total of 34,966 peaks called in large size foetal piglets 9.82% (3436 peaks), and from 49,060 in small size foetal piglets 12.2% (6031 peaks), were specific to each size group. A detailed summary of the number of peaks for each size is included in Supplementary File 1. The read count frequency around TSSs of the size specific sets of peaks was noticeably lower than when all peaks (i.e. including both those that were shared and those that were size

specific) were analysed (Figure 5A). The results show both a shared chromatin openness profile and size specific openness profiles indicating a difference in genome regulation between the three sizes of foetal piglet at Day 90 of gestation (Figure 5A). The distribution of ATAC-Seq peaks within genomic features were skewed towards intergenic regions for the large and small sized foetal piglets as shown in Figure 5B. Fewer size-specific ATAC-Seq peaks were located within 1Kbp of TSSs in comparison with the ATAC-Seq peaks that were shared across all three sizes of foetal piglet (Figure 5C).

Figure 5: Mapping of ATAC-Seq peaks at Day 90 of gestation that were: i) specific to small or large sized piglets or ii) shared across samples. A) mapping relative to the TSS, B) distribution within genomic features and C) distribution of relative to the TSS.

Analysis of gene expression using RNA-Seq

We generated RNA-Seq data from the same muscle tissue samples that were used to generate the ATAC-Seq libraries, in order to link regions of open chromatin with gene expression. The transcript expression estimates for the muscle tissue samples from the five developmental time points (26 samples in total) were calculated as Transcript per Million mapped reads (TPM) using Kallisto. The TPM expression estimates were then investigated using PCA (Figure 6) to identify any samples that did not group as expected according to developmental time point. The majority of the samples from each developmental time point clustered together in the first two dimensions of the PCA (Figure 6). The only exception was sample D90S121016 which we excluded from the day 90 small sized foetal piglet group based on the proximity of its expression profile to the Day 45 samples (Figure 6). We believe this

was an issue related to a mix up with the tissue samples, which led us to extract a day 45 sample by mistake, as the ATAC-Seq library for the corresponding sample had clustered as expected with the other day 90 samples (Supplementary Figure S2).

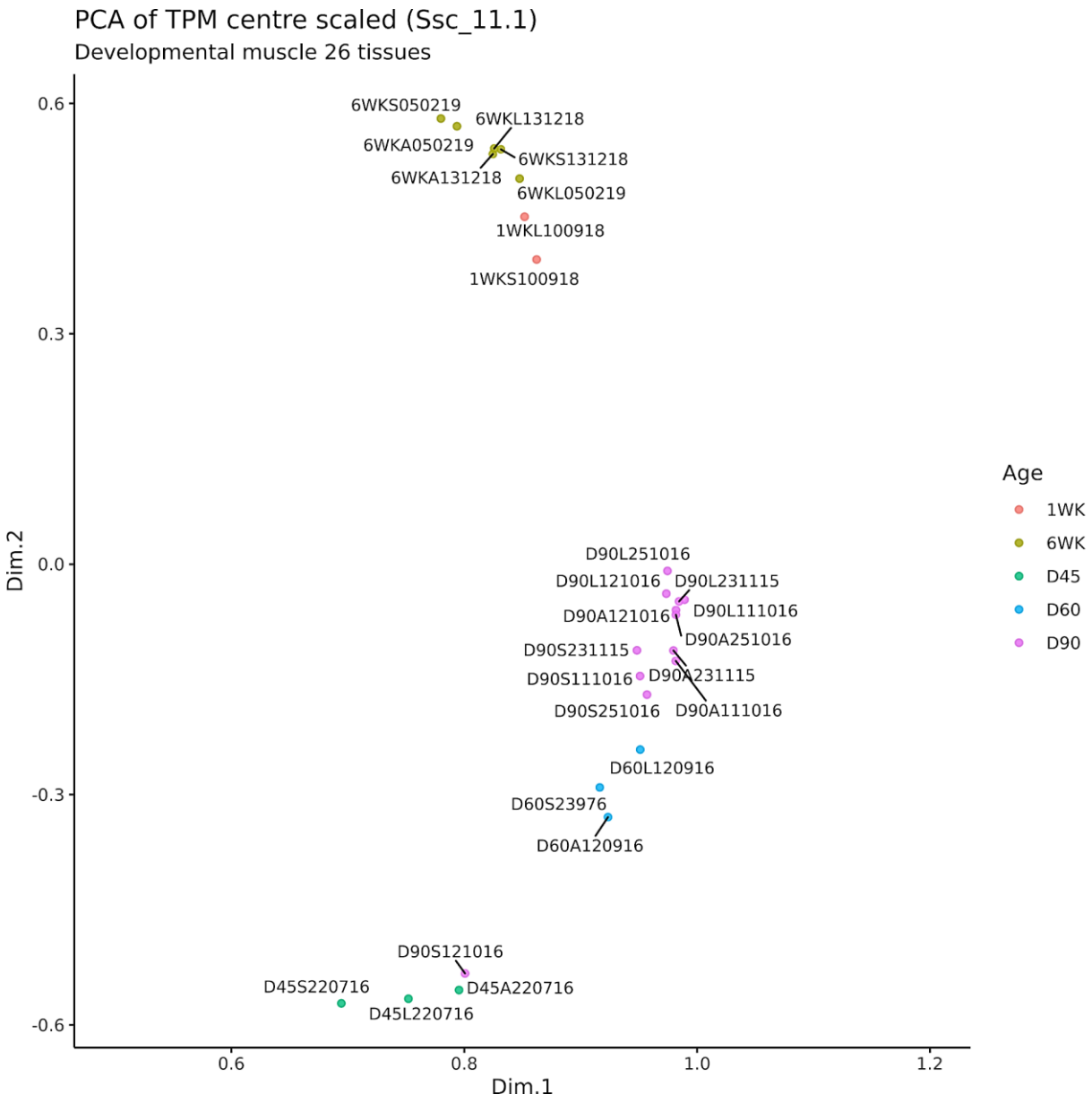


Figure 6: Principal Component Analysis (PCA) of gene expression estimates (as TPM) from the RNA-Seq data for each sample. The samples cluster according to developmental stage with clear separation of neonatal and post-natal samples. D45 = Gestational Day 45; D60 = Gestational Day 60; D90 = Gestational Day 90; 1WK = Neonatal 1 week old ;6WK = Juvenile 6 weeks old.

Analysis of genes that were differentially expressed between the three sizes of foetal piglet at Day 90 of gestation

Differential gene expression analysis was performed using the TPM values for the three sizes of foetal piglet (small, average, large) at Day 90 of gestation. Between the three sizes of foetal piglet 89 genes (FDR 10%), were found to be differentially expressed. When average vs small sized foetal piglets were compared, 58 up- and 31 down-regulated genes were detected. When large vs small sized foetal piglets were compared, 54 up- and 35 down-regulated genes were detected. Differentially expressed genes with an adjusted *p* value (FDR < 0.1) and log2 fold change (log2FC) >= 0.1 are annotated in Figure 7. The comparison between large and small sized foetal piglets resulted in the largest number (n=89) of differentially expressed genes. The list of differentially expressed genes and detailed analysis metrics can be found in Supplementary Table 5. Many of the genes that were differentially expressed between large and small and average and small foetal piglets are involved in skeletal muscle function and growth (Figure 7). The gene calsequestrin 1 (*CASQ1*), for example, which was 1.54 fold up-regulated (log2FC 0.63 ± 0.17 adjusted *p* value = 2.0e-02) in large relative to small foetal piglets is the skeletal muscle specific member of the calsequestrin protein family, and is highly expressed in skeletal muscle in adult pigs, see (http://biogps.org/pigatlas/) (Freeman et al., 2012; Summers et al., 2020). *MYBPC2*, a gene that encodes myosin binding protein C, was also 2 fold up-regulated (log2FC 1.02 ± 0.22 adjusted *p* value =3.78e-04) in large relative to small foetal piglets (Figure 7). It has also been shown to be up-regulated in the muscle of adult pigs, see (http://biogps.org/pigatlas/) (Freeman et al., 2012; Summers et al., 2020). The muscle specific transcription factor myogenin (*MYOG*) was down-regulated, (log2FC 0.28 ± 0.09 adjusted *p* value = 9.0e-02), in small relative to large foetal piglets (Figure 7).

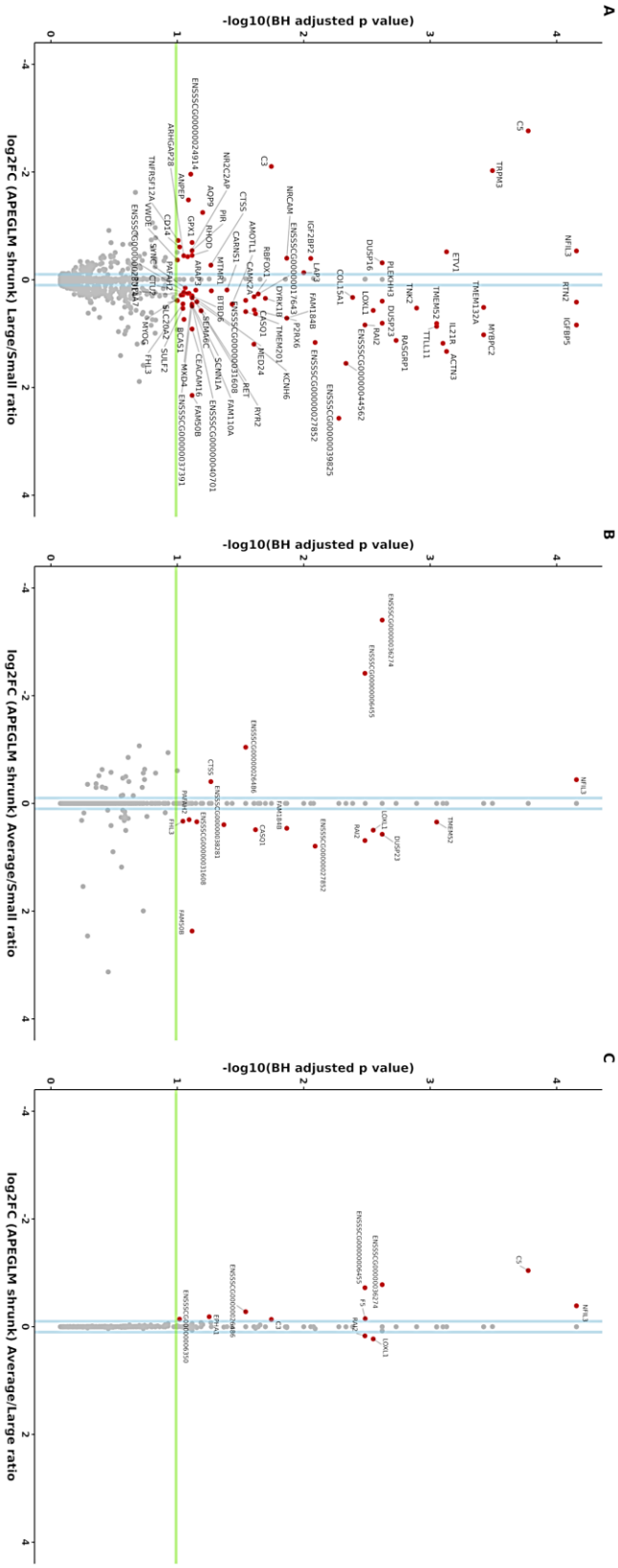


Figure 7: Differentially expressed genes between large, average and small sized piglets at Day 90 of gestation. A) Large vs Small B) Average vs Small C) Average vs Large. Differentially expressed genes are shown in red, the log2FC >0.1 in blue and the significance threshold as

a green line. The ApeGLM shrinkage method was used to normalised the log fold change plotted on the x axis. Detailed methodology is described in (Love et al., 2014).

Overlay of the RNA-Seq differentially expressed genes and ATAC-Seq peaks from large vs small foetal piglets at Day 90 of gestation

A further overlay of the ATAC-Seq and RNA-Seq datasets was performed for the Day 90 large and small sized foetal piglets. ATAC-Seq peaks annotated using the *Sscrofa11.1* Ensembl gene track information (black) and differentially expressed genes between the large vs small sized foetal piglets at Day 90 (green) are shown in Figure 8. This analysis allowed us to determine which of the differentially expressed genes had an ATAC-Seq peak that was specific to either large or small sized piglets in its vicinity. The distribution of ATAC-Seq peaks around TSSs (within a 3Kbp distance) was plotted for peaks specific to the large foetal piglets (Figure 8A), or specific to the small foetal piglets (Figure 8B). Size specific peaks within the 5'UTR region of four differentially expressed genes, *MYOG,* ryanodine receptor 2 (*RYR2*), transmembrane 4 L six family member 4 (*TM4SF4*) and interleukin 21 receptor (*IL21R*) (Figure 7), were only observed in the small foetal piglets (Figure 8B). There was no evidence of size specific peaks near these genes in the large sized foetal piglets (Figure 8B). Of the four genes, *MYOG* is known to be highly expressed in skeletal muscle tissue, see (http://biogps.org/pigatlas/) (Freeman et al., 2012; Summers et al., 2020). In some cases presence of a size-specific ATAC-Seq peak within the 5' UTR region was associated with down-regulation of genes involved in muscle growth in small relative to large piglets. *MYOG*, for example, was down-regulated in small sized foetal piglets (Figure 7), which coincided with a regulatory region 315 bp in size, 8769 bp upstream of the TSS, that was present in the small sized piglets but absent in the large sized foetal piglets (Figure 8 A&B).

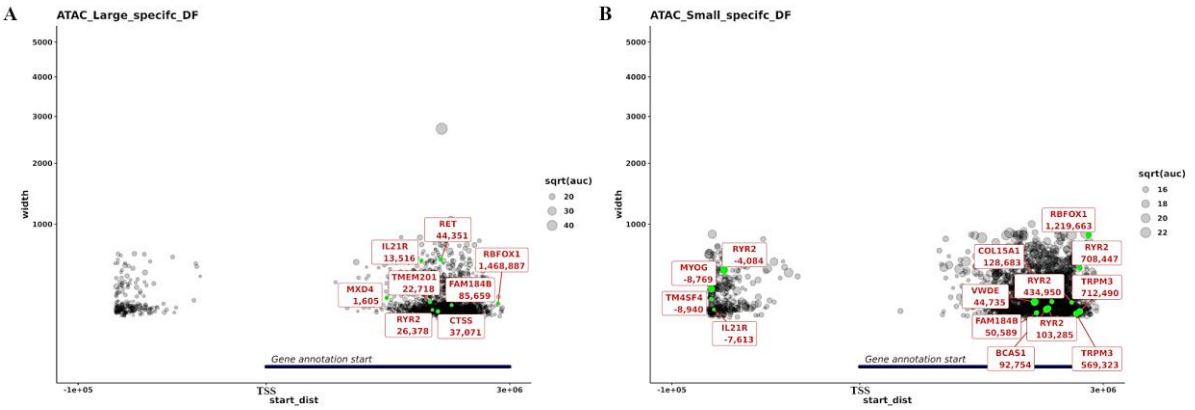


Figure 8: Proximity of ATAC-Seq peaks specific to large (A) and small (B) piglets and differentially expressed genes. Differentially expressed genes are marked in green. The x-axis (start_dist) is the distance from the start of the gene model to the start of the ATAC-Seq peak, for +ve values the peak is either within the gene or within 10kbp of the 3' end and for -ve

values the peak is within 10Kbp of the 5' end of the gene. The y-axis indicates the width of the peak. As the y axis represent the width of the peak, the larger the node the wider the ATAC-Seq peak.

Association between transcript expression and chromatin openness

Differentially expressed genes showed a significant number of size specific peaks in small compared to large piglets (*p* value = 0.007). Although this difference did not directly translate to a difference in DNA openness score (DOS) between the sizes of piglet. A DOS was calculated for every transcript to correlate transcript expression (as TPM) with chromatin openness in each of the small and large sized foetal piglets separately at Day 90 of gestation. In large sized foetal piglets (Figure 9 B, D & F) the transcript expression was significantly higher (*p* value=0.006) on average than in the small sized foetal piglets (Figure 9 A, C & E). However, there was no overall trend present between the transcript TPM and DOS values in the large or small sized foetal piglets. This observation was made based on the lack of any overlap between the top 30 highly expressed transcripts (annotated by transcript IDs in Figure 9) between the large and small sized foetal piglets. Clustering the data set, based on transcript expression and DOS did not result in any meaningful grouping either (shown as four colours of dots in Figure 9).

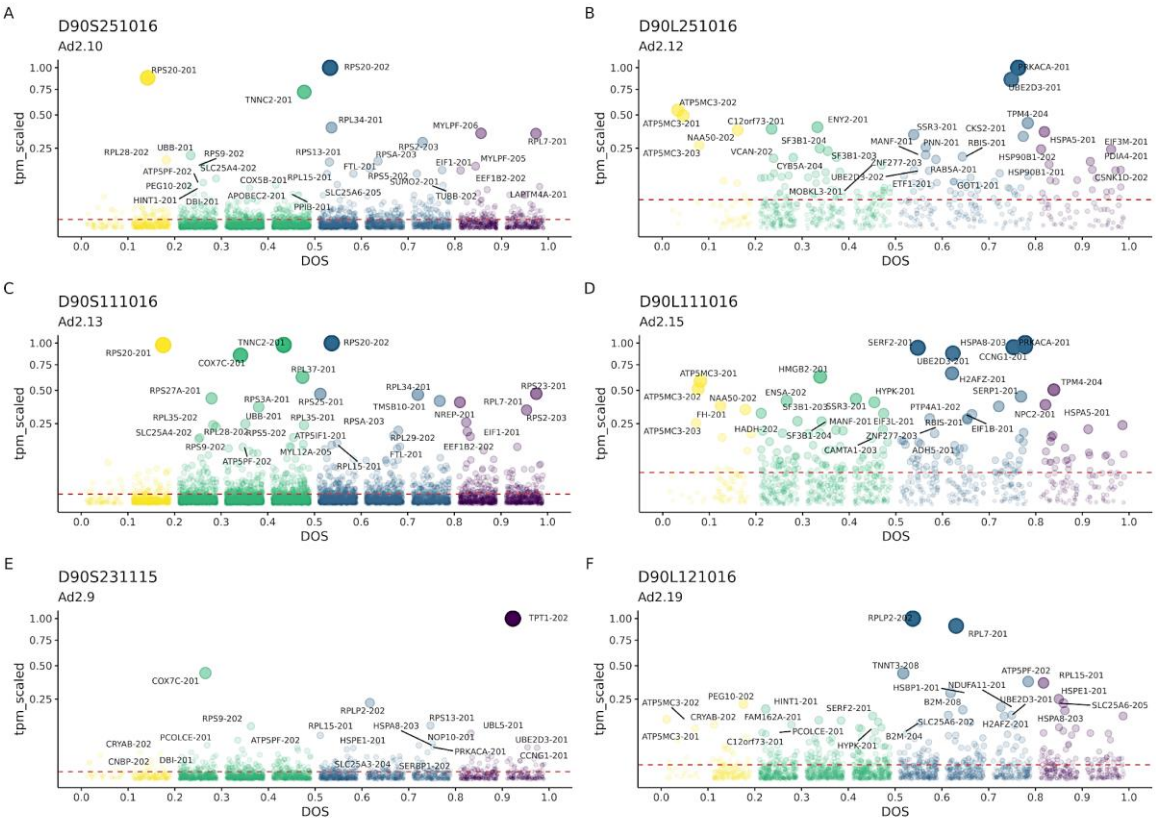


Figure 9: The overlay of DNA openness score (DOS) and transcript expression level (as TPM) in large and small sized piglets from three litters: A & B) Small and large piglets from litter

251016. C & D) small and large piglets from litter 111016. E) small piglet from 231115 (no RNA-Seq sample from a large piglet was available for this litter from comparison) and F) large piglet from 121016 (no RNA-Seq sample from a small piglet was available for this litter from comparison). Dots are coloured and sized according to increasing TPM. Colours represent DOS binned (10 bins) categories of 0.8-1 (purple), 0.5-0.8 (blue), 0.2-0.5 (green) and 0-0.2 in yellow. The average (mean) expression of transcripts in each sample marked by a red dotted line.

Allele specific expression analysis of RNA-Seq data and overlap with size specific ATAC-Seq peaks

An imbalance in allele specific expression (ASE) was present in 17 genes expressed in foetal piglets at day 90 of gestation. Four genes *PXDN,* Collagen Type I Alpha 1 Chain (*COL1A1*), DEAD-box helicase 6 (*DDX6*) and *ENSSSCG00000009327 (*High Mobility Group Box 1, *HMGB1),* showed significant ASE in all three sizes of piglet (Figure 10A). Peroxidasin (*PXDN*) exhibited significant allelic imbalance in the small sized foetal piglets, with 100% representation across all three biological replicates (Figure 10A). Peroxidasin (*PXDN*) is a basement membrane protein involved in tissue integrity and formation of extracellular matrix (Colon et al., 2017). *COL1A1*, which is also involved in formation of the extracellular matrix exhibited significant allelic imbalance in all three biological replicates of both large and small foetal piglets (100% representation) (Figure 10A). *ARRDC3,* which encodes arrestin domain containing 3, on the other hand exhibited allelic imbalance only in the small sized foetal piglets, across all three biological replicates (100% representation). The ATAC-Seq peaks specific to the small sized foetal piglets associated with genes exhibiting allelic imbalance are visualised in Figure 10B. The ATAC-Seq peaks associated with *PXDN* and *COL1A1* were present in both the large and small sized foetal piglets but at different genomic coordinates specific to each size (Figure 10B). The ATAC-Seq peaks associated with *ARRDC3* were only present in the small sized foetal piglets (Figure 10B).

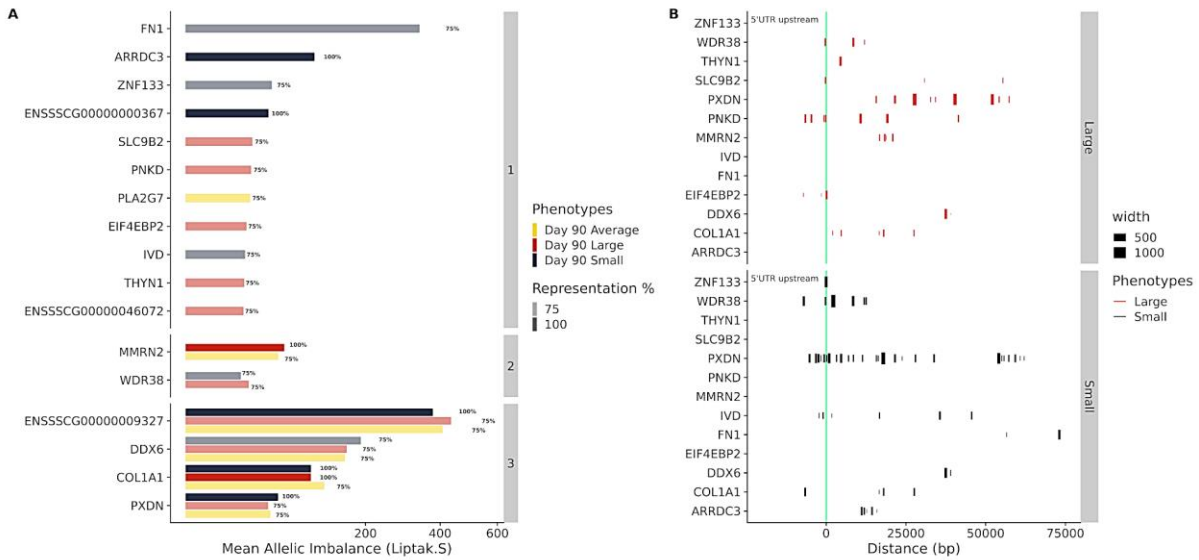


Figure 10: Genes exhibiting significant allele specific expression in small, average and large sized piglets at Day 90 of gestation. A) genes exhibiting allelic imbalance in up to three biological replicates for each size of piglet (genes were classified into three groups according to whether allelic balance was detected in all three biological replicates: 3= exhibited in large, average and small, 2= exhibited in either small and large or average and large, 1= exhibited only in one group). The percentage representation (75% to 100 %) of the gene exhibiting allelic imbalance across biological replicates is indicated by the transparency of the bars. The genes showing ASE in each of the three sizes of piglet are indicated using coloured bars that represent the level of allelic imbalance calculated using the Liptak score (and output metric of the GeneiASE pipeline). B) Distance of ATAC-Seq peaks, specific to either large or small sized piglets, from the TSS of genes showing significant allelic imbalance.

Allele-specific expression positive variants in small sized piglet haplotypes

A read backed approach was used for haplotype phasing the RNA-Seq data from the small sized foetal piglets. This haplotype block discovery method was performed to locate potential variants (and corresponding genotypes) that might contribute to the allelic imbalance in expression observed in the genes described above. Overall, 53,439 haplotype blocks were identified in three biological replicates with an average length of 379 bp ± 2,881bp (mean ± SD). The longest block was located on chromosome 14 and was 248kbp length (chr14:56857698-57105958) overlapping the first two exons of solute carrier family 35 member F3 (*SLC35F3*) and the complete genomic coordinates of *U6*. Haplotype blocks residing within coordinates of multiple ATAC-Seq peaks in genes showing ASE (i.e. *ARRDC3* and *PXDN)* are shown using IGV genomic browser in Figures 11 and 12. Two haplotype blocks (corresponding IDs 98441748 and 98442176 in Supplementary File 5; *D90S_phased.gtf*) in the first intronic region of *ARRDC3* overlapped with ATAC-Seq peaks

specific to small sized foetal piglets shown in pink and black regions in Figure 11. Two haplotypes blocks containing variants exhibiting imbalance in allelic expression were also identified (haplotype block ID 98440164 containing variant *rs320066059* and ID 98444537 containing variant *rs343363588*) in the second and first exonic regions of *ARRDC3* respectively. A similar overlay of the genomic coordinates of *PXDN* showed 11 ASE positive variants overlapping four haplotype blocks (IDs 132337596, 132349428, 132357889 and 132361644) (Figure 12). These haplotype blocks span across multiple exons of four known isoforms of *PXDN* (exons 6 – 23), shown as green blocks in Figure 12. Two haplotype blocks (IDs 26618720 and 26619492) containing three ASE positive variants (*rs332129500*, *rs336409927* and *rs320640601*) were also detected in the *ZNF133* gene (Supplementary Figure S5). Supplementary File 5 contains all the tracks required for visualisation of this analysis in the Integrated Genomic Viewer (IGV) (Thorvaldsdóttir et al., 2013).

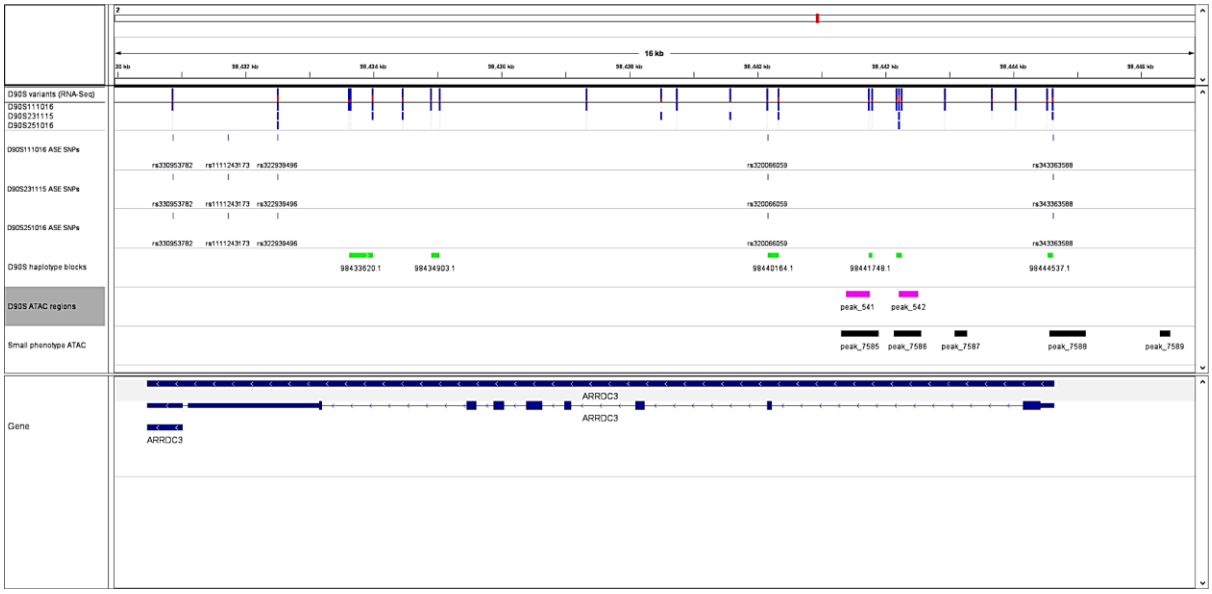


Figure 11: Genomic track visualisation of the genomic regions including gene ARRDC3. Variants called using RNA-Seq data with read-backed haplotype blocks in the small sized piglets, and the corresponding open chromatin regions (ATAC-Seq peaks), shown. The top four tracks show variants in dark blue (one consensus genotype and three allele-specific positive variants in each piglet) and the haplotype block in green. ATAC-Seq peaks specific to small sized piglets are shown in pink and shared ATAC-Seq peaks in black.

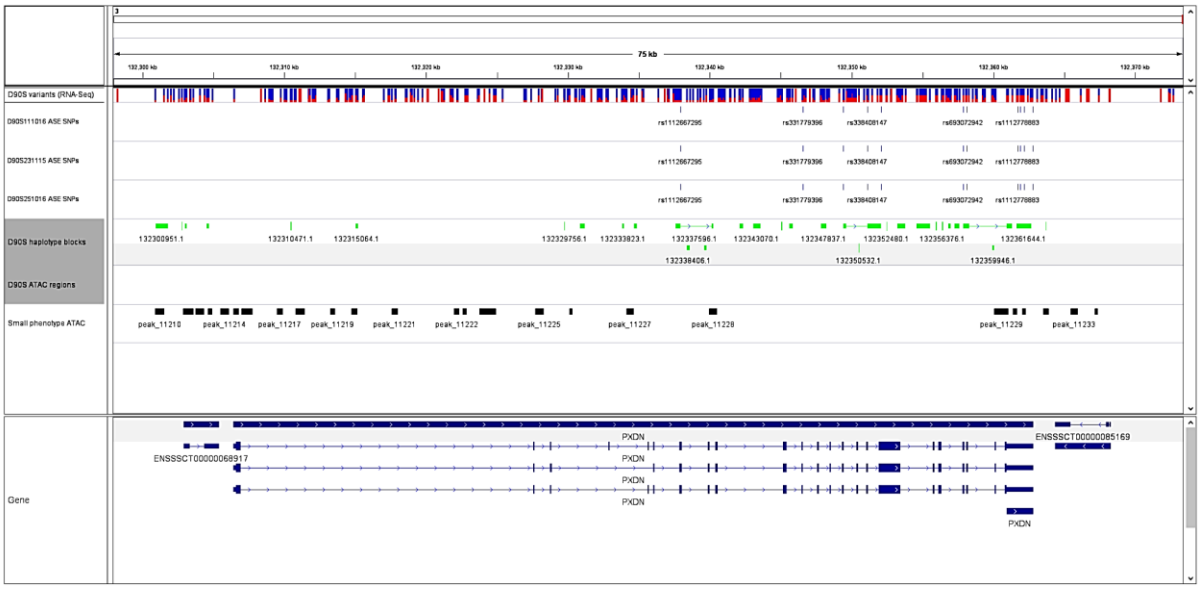


Figure 12: Genomic track visualisation of the genomic region including gene PXDN. Variants called using RNA-Seq data with read-backed haplotype blocks in the small sized piglets, and the corresponding open chromatin regions (ATAC-Seq peaks), are shown. The top four tracks show variants in dark blue (one consensus genotype track and three separate tracks for allele-specific positive variants in each of the three piglets) and the haplotype block in green. ATAC-Seq peaks specific to small sized piglets are shown in pink and shared ATAC-Seq peaks in black.

Discussion

In this study we used ATAC-Seq and RNA-Seq to improve our understanding of gene expression and regulation in developing pig muscle. The aims of the study were to: 1) Optimise the Omni-ATAC-Seq protocol for frozen pig muscle tissue; 2) Map regions of open chromatin in semitendinosus muscle tissue from small, average and large sized male piglets at five developmental stages (Day 45, 60 & 90 of gestation, one and six weeks post-natal) and 3) Integrate ATAC-Seq peaks with RNA-Seq data to link regions of open chromatin with gene and allele specific expression information.

To fulfil aim one, we optimised the Omni-ATAC-Seq protocol (Corces et al., 2017) for frozen muscle tissue. This to our knowledge is the first time the Omni-ATAC-Seq protocol (Corces et al., 2017) has been optimised for frozen tissue from a farmed animal species. Other studies have used ATAC-Seq to profile open chromatin in freshly sorted cell types e.g. (Foissac et al., 2019) or isolated and cryopreserved nuclei from dissociated tissue e.g. (Halstead et al., 2020b, 2020a). Working with sorted cells was outside the scope of this study. We were, however, able to perform a comparison of ATAC-Seq libraries prepared from either flash frozen tissue or cryopreserved nuclei for a small subset of samples. We found that the

datasets generated by the two methods were broadly comparable, in terms of the distribution of ATAC-Seq reads mapping to genomic regions and in the percentage of reads mapping to exons. The ability to utilise flash frozen tissue effectively for ATAC-Seq is advantageous for two reasons. Firstly, many legacy samples from large animal studies have been flash frozen then archived at -80°C and represent a very valuable resource if they can be utilised. Secondly, flash freezing is straightforward to undertake and standardise, especially when the logistics of collecting samples from large animals can be technically challenging (Wong et al., 2012). For this study we have only optimised the Omni-ATAC-Seq protocol for flash frozen muscle tissue samples from pigs. Expanding the protocol to other tissues and other species should be relatively straightforward, although some tissue-specific optimisation will be required, particularly for tissues that are known to be complex to work with.

For aim two of the study, we generated open chromatin profiles, in the form of ATAC-Seq peaks, for semitendinosus muscle from piglets from five developmental stages. The developmental stages were chosen according to their relevance to muscle development. ATAC-Seq peaks mapped as expected to promotor regions and within 1 Kbp of the TSS. The width of the ATAC-Seq peaks, and therefore the openness of chromatin, differed throughout development (Detailed BED files can be found in Supplementary file 1). The widest peaks were detected at Day 90 of gestation. The distribution of ATAC-Seq peaks in intergenic regions at Day 90 in large and small sized piglets indicated that piglets of different sizes demonstrate changes in genome regulation primarily at intergenic sites (i.e. differential enhancer activity). Day 90 is a critical stage of muscle development when fibre formation ceases and muscle growth accelerates through fibre hypertrophy (Oksbjerg et al., 2004). Significant up-regulation of genes involved in muscle growth also occurs at Day 90 (Zhao et al., 2015; Ayuso et al., 2016). As such chromatin may be more open at this developmental stage to allow transcription factor binding prior to the rapid muscle growth that occurs during the early postnatal period (Rudar et al., 2019). The read count frequency around the TSS also differed throughout development, increasing to Day 90 then decreasing postnatally. This trend indicates that the global open chromatin profile of muscle tissue changes during gestation in pigs. Human studies profiling open chromatin in preimplantation embryos found global differences in chromatin accessibility between embryo stages (Wu et al., 2018; Liu et al., 2019; Halstead et al., 2020c). ATAC-Seq datasets for post implantation embryos in humans and other mammalian species are limited. ChIP-seq analyses of a wide variety of histone markers in the brain, heart, and liver of early human embryos identified developmental stage-specific patterns in the epigenome (Yan et al., 2016). The sample size in our study was small, with only a few biological replicates for most points, with the exception of Day 90. Even so, our results suggest that chromatin accessibility and regulation of gene expression changes throughout development in the muscle of production pigs. This is significant for studies aiming

to understand when during development functional variation in the genome has an effect on the adult phenotype.

To address aim three, we overlapped ATAC-Seq peaks with RNA-Seq data to link regions of open chromatin with gene and allele specific expression information. We used this approach to compare the gene expression and chromatin openness between foetal piglets of different size at Day 90 of gestation. Analysis of the ATAC-Seq data indicated that at Day 90 of gestation ATAC-Seq peaks were narrower, and therefore chromatin was less open, in small foetal piglets and wider, and therefore more open in large foetal piglets. This finding, though the sample size for the study is small, is indicative of a difference in genome regulation between large and small foetal piglets. Other studies have used a similar approach to investigate the effect of histone modification on the expression of genes involved in placental development in pigs (Han et al., 2019). Differences in open chromatin were reflected in the expression of genes involved in muscle growth. Analysis of the RNA-Seq data revealed that genes associated with muscle growth, including *CASQ1, MYBPC2* and *MYOG*, were differentially expressed in large relative to small piglets. *CASQ1* encodes the skeletal muscle specific member of the calsequestrin protein family, is related to muscle metabolism, and has been shown to be highly expressed in fat pig breeds (Zhao et al., 2011). In this study, *CASQ1* was up-regulated in large relative to small foetal piglets. *MYBPC2* encodes the fast isoform of the myosin binding protein C family (Weber et al., 1993). In Piedmontese (*GDF8* mutant) cattle *MYBPC2* is highly expressed in foetal muscle, reflecting fast glycolytic fibre structural differentiation (Lehnert et al., 2007). In this study, *MYBPC2* was highly up-regulated in large versus small foetal piglets, potentially reflecting a greater proportion of fast glycolytic muscle fibres.

MYOG was down regulated in small sized foetal piglets relative to large sized foetal piglets. *MYOG* is essential for myoblast fusion during muscle development (https://www.uniprot.org/uniprot/P49812) and associated with QTLs, for body weight at birth (https://www.animalgenome.org/cgi-bin/QTLdb/SS/qdetails?QTL_ID=8656) and backfat thickness (https://www.animalgenome.org/cgi-bin/QTLdb/SS/qdetails?QTL_ID=8657), according to Genome Wide Association Studies (GWAS) (Xue and Zhou, 2006). Other studies measuring gene expression also found *MYOG* was down-regulated in muscle cell types from low birth weight piglets (Felicioni et al., 2020; Stange et al., 2020) and in pigs with high levels of intramuscular fat (Lim et al., 2017). When we compared the ATAC-Seq and RNA-Seq data we identified an ATAC-Seq peak within the 5' UTR region of *MYOG* (315bp in size, 8769bp upstream of the TSS) that was present in the small foetal piglets but missing from the large foetal piglets for Day 90 of gestation. This peak directly correlated with down-regulation of *MYOG* in small relative to large foetal piglets. In future work we plan to remove this peak using CRISPR genome editing and measure the effect on primary muscle cells in culture. Further

validation of the variants within this regulatory region would be useful to determine whether they might underlie variation in intramuscular fat or myofiber specification as was the case for the regulatory variant recently characterised in *MYH3* (Cho et al., 2019).

Further analysis of the RNA-Seq data revealed several genes exhibiting allelic expression imbalance. This imbalance of expression could be linked to ATAC-Seq size-specific peaks present only in the small foetal piglets at Day 90 of gestation. Size-specific ATAC-Seq peaks associated with allelic expression imbalance in the gene *ARRDC3,* for example, were only present in the small sized foetal piglets. *ARRDC3* is associated with obesity and adipocyte differentiation and function in mice (Carroll et al., 2017) and with growth and muscularity traits in cattle (Bolormaa et al., 2014; Saatchi et al., 2014; Abo-Ismail et al., 2017; Carroll et al., 2017). Allelic imbalance and a corresponding ATAC-Seq peak might be indicative of dysregulation of this gene in muscle and fat tissue formation in the small sized piglets. We used haplotype phasing to identify candidate genetic variants that might be causing the allelic imbalance in *ARRDC3* expression. The haplotype blocks were consistent across all three small sized foetal piglets. These variants in *ARRDC3,* and those within other genes exhibiting an imbalance in expression, including *PXDN* and *ZNF133,* would be suitable candidates for further functional validation. Variants with this type of functional information have been shown in cattle to be useful for the prediction of complex traits (Xiang et al., 2019).

The datasets we have generated for this study lay a foundation for incorporating functional information in statistical analyses, to increase the precision and power with which we can fine map high quality causal variants in pigs. This would make it possible to increase the accuracy of genomic selection and the efficiency with which breeding turns genetic variation into genetic gain. The next stage of the study is to leverage the ATAC-Seq and RNA-Seq data with a very large dataset of genetic variants from production pigs to determine whether any trait-linked variants are located within the open chromatin regions we have identified for muscle tissue. The characterisation of regulatory and expressed regions of the genome in muscle tissues also provides a basis for genome editing to promote functional genomic variants in pig breeding programmes (Jenko et al., 2015; Hickey et al., 2016; Johnsson et al., 2019), providing a route to application for FAANG data.

Conclusions

The dataset we have generated provides a powerful foundation to investigate how the genome is regulated in production pigs and contributes valuable functional annotation information to define and predict the effects of genetic variants in pig breeding programmes. The outcomes of the study will: 1) help us to understand the molecular drivers of muscle growth in pigs; 2) provide a foundation for functionally validating target genomic regions in *in*

vitro systems and 3) identify high quality causative variants for muscle mass with the goal of harnessing genetic variation and turning it into sustainable genetic gain in pig breeding programmes.

Acknowledgements

The authors would like to thank the staff and farm technicians at both Dryden Farm and Easter Howgate. Charis Hogg provided assistance in planning and collection of samples from the foetal time points and Christopher Proudfoot coordinated sampling of the six-week-old pigs. ATAC-Seq and RNA-Seq libraries were sequenced by Edinburgh Genomics. The authors are also grateful for the support of the FAANG Data Coordination Centre (http://data.faang.org) in the upload and archiving of the sample data and metadata.

Ethics Statement

All projects from which samples for this study were collected were reviewed and approved by The Roslin Institute, University of Edinburgh's Animal Work and Ethics Review Board (AWERB). All animal work was carried out under the regulations of the Animals (Scientific Procedures) Act 1986.

Data Availability

The raw sequence data for the ATAC-Seq samples (n=24) are available via the European Nucleotide Archive (ENA) under accession number PRJEB41485. Details of all samples processed for the RNA-Seq dataset (n=26) can be accessed via the ENA under accession number PRJEB41488. The sample metadata is available via the BioSamples database under sample accession numbers SAMEA7178119, SAMEA7178120, SAMEA7178122, SAMEA7178123, SAMEA7178124, SAMEA7178125, SAMEA7178126, SAMEA7178127, SAMEA7178134, SAMEA7178138, SAMEA7178142, SAMEA7178149, SAMEA7178150, SAMEA7178153, SAMEA7178159, SAMEA7178160, SAMEA7178164, SAMEA7178178, SAMEA7178179, SAMEA7178180, SAMEA7178182, SAMEA7178183, SAMEA7178184, SAMEA7178185, SAMEA7178187 and SAMEA7178188. These datasets are curated and submitted to FAANG data portal according to FAANG's sample and experimental guidelines (Harrison et al., 2018). All the sample, experiment and analysis protocols for this study are also available through the FAANG Data Coordination Centre via the following links: https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_ATAC_Seq_DNAIsolationandTagmentation_Frozen_Muscle_Tissue_20200720.pdf , https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_ATAC-Seq_DNAIsolationandTagmentation_Cryopreserved_Muscle_Nuclei_Preparations_2020072

0.pdf , https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_ATAC-Seq_LibraryPreparationandSizeSelection_20200720.pdf , https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_RNA_IsolationoftotalRNAfromfrozentissuesamples_20200720.pdf , https://data.faang.org/api/fire_api/experiments/ROSLIN_SOP_ATAC-Seq_Cryopreservednucleisamplesfrompigmuscletissue_Experimental_Protocol_20200720.pdf , https://data.faang.org/api/fire_api/experiments/ROSLIN_SOP_ATAC-Seq_Frozenpigmuscletissuesamples_Experimental_Protocol_20200720.pdf , https://data.faang.org/api/fire_api/experiments/ROSLIN_SOP_RNA_RNAIsolationandSequencingofPigMuscleTissues_Experimental_Protocol_20200720.pdf , https://data.faang.org/api/fire_api/analyses/ROSLIN_SOP_ATAC-Seq_analysis_pipeline_20201113.pdf , https://data.faang.org/api/fire_api/analyses/ROSLIN_SOP_RNA-Seq_analysis_pipeline_20201113.pdf , https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_Collection_of_tissue_samples_for_ATAC-Seq_and_RNA-Seq_from_large_animals_20200618.pdf , https://data.faang.org/api/fire_api/samples/ROSLIN_SOP_Cryopreservation_of_Nuclei_for_ATACSeq_using_GentleMACS_20201119.pdf**.** All the supplementary tables, figures and files are also available from https://doi.org/10.6084/m9.figshare.13562285.

Author Contributions

ELC devised the study and acquired the funding with MAH, FXD and ALA. CJA and CS designed the experiment from which the foetal tissues were collected. YCA, FXD, CS and CJA collected the samples from the foetal and one week old piglets. MS and ELC collected the samples from the six week old piglets. MS and ELC performed the cryopreservation of nuclei. ELC and MAH performed the ATAC-Seq library preparation. ELC and YCA extracted the RNA. MMH provided advice on ATAC-Seq library preparation and data analysis. SAW collated all sample metadata for the project. MS performed all bioinformatic and data analysis, except the analysis of ASE which he performed with SAW. ELC wrote the manuscript with MS. MJ provided critical assessment of the manuscript. All authors read and approved the final version.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Code Availability

The bioinformatic pipelines used for processing the ATAC-Seq (mapping and peak calling), RNA-Seq (transcript level expression analysis) and Allele specific expression are available via a code repository at https://msalavat@bitbucket.org/msalavat/pig_muscle.git respectively (https://bitbucket.org/msalavat/pig_muscle/src/master/).

Funding

The work was funded by an Institute Strategic Programme Pump-Priming grant ‘ Profiling Open Chromatin in Developing Pig Muscle’, BBSRC Institute Strategic Program grants awarded to the Roslin Institute ‘Farm Animal Genomics’ (BBS/E/D/2021550) and ‘Prediction of genes and regulatory elements in farm animal genomes’ (BBS/E/D/10002070) and by BBSRC grant ‘Ensembl – adding value to animal genomes through high quality annotation’ (BB/S02008X/1) . ELC and MAH are both supported by University of Edinburgh Chancellors’ Fellowships. This research was also funded in part by the Bill & Melinda Gates Foundation and with UK aid from the UK Foreign, Commonwealth and Development Office (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland’s Rural College), and the International Livestock Research Institute. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation nor the UK Government.

References

Abo-Ismail, M. K., Brito, L. F., Miller, S. P., Sargolzaei, M., Grossi, D. A., Moore, S. S., et al. (2017). Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. *Genet. Sel. Evol.* 49, 82. doi:10.1186/s12711-017-0356-8.

Aiello, D., Patel, K., and Lasagna, E. (2018). The myostatin gene: an overview of mechanisms of action and its relevance to livestock animals. *Anim. Genet.* 49, 505–519. doi:10.1111/age.12696.

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16, 57. doi:10.1186/s13059-015-0622-4.

Ashmore, C. R., Addis, P. B., and Doerr, L. (1973). Development of Muscle Fibers in the Fetal Pig. *J. Anim. Sci.* 36, 1088–1093. doi:10.2527/jas1973.3661088x.

Ayuso, M., Fernández, A., Núñez, Y., Benítez, R., Isabel, B., Fernández, A. I., et al. (2016).

Developmental stage, muscle and genetic type modify muscle transcriptome in pigs: Effects on gene expression and regulatory factors involved in growth and metabolism. *PLoS One* 11, e0167858. doi:10.1371/journal.pone.0167858.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR. *Source J. R. Stat. Soc. Ser. B* 57, 289–300. Available at: https://www.jstor.org/stable/pdf/2346101.pdf?refreqid=excelsior%3A8080132077b9217878a58e47abbcf2d8 [Accessed October 27, 2015].

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–20. doi:10.1093/bioinformatics/btu170.

Bolormaa, S., Pryce, J. E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., et al. (2014). A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. *PLoS Genet.* 10, 1004198. doi:10.1371/journal.pgen.1004198.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi:10.1038/nbt.3519.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* 109, 21.29.1-21.29.9. doi:10.1002/0471142727.mb2129s109.

Carroll, S. H., Zhang, E., Wang, B. F., LeClair, K. B., Rahman, A., Cohen, D. E., et al. (2017). Adipocyte arrestin domain-containing 3 protein (Arrdc3) regulates uncoupling protein 1 (Ucp1) expression in white adipose independently of canonical changes in β-adrenergic receptor signaling. *PLoS One* 12. doi:10.1371/journal.pone.0173823.

Cho, I.-C., Park, H.-B., Ahn, J. S., Han, S.-H., Lee, J.-B., Lim, H.-T., et al. (2019). A functional regulatory variant of MYH3 influences muscle fiber-type composition and intramuscular fat content in pigs. *PLOS Genet.* 15, e1008279.

Clark, E. L., Archibald, A. L., Daetwyler, H. D., Groenen, M. A., Harrison, P. W., Houston, R. D., et al. (2020). From FAANG to Fork: Application of Highly Annotated Genomes to Improve Farmed Animal Production. (doi: 10.20944/preprints202010.0118.v1). *Preprints*. doi:10.20944/preprints202010.0118.v1.

Colon, S., Page-Mccaw, P., and Bhave, G. (2017). Role of hypohalous acids in basement membrane homeostasis. *Antioxidants Redox Signal.* 27, 839–854. doi:10.1089/ars.2017.7245.

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* 14, 959–962. doi:10.1038/nmeth.4396.

Edinburgh, U. of (2020). Edinburgh Compute and Data Facility. Available at: https://www.ed.ac.uk/is/research-computing-service [Accessed July 6, 2020].

Edsgärd, D., Iglesias, M. J., Reilly, S.-J., Hamsten, A., Tornvall, P., Odeberg, J., et al. (2016). GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci. Rep.* 6, 21134. doi:10.1038/srep21134.

Estany, J., Ros-Freixedes, R., Tor, M., and Pena, R. N. (2017). TRIENNIAL GROWTH AND DEVELOPMENT SYMPOSIUM: Genetics and breeding for intramuscular fat and oleic acid content in pigs1. *J. Anim. Sci.* 95, 2261–2271. doi:10.2527/jas.2016.1108.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi:10.1093/bioinformatics/btw354.

Felicioni, F., Pereira, A. D., Caldeira-Brant, A. L., Santos, T. G., Paula, T. M. D. D., Magnabosco, D., et al. (2020). Postnatal development of skeletal muscle in pigs with intrauterine growth restriction: morphofunctional phenotype and molecular mechanisms. *J. Anat.*

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17, 108. doi:10.1186/s12915-019-0726-5.

Freeman, T. C., Ivens, A., Baillie, J. K., Beraldi, D., Barnett, M. W., Dorward, D., et al. (2012). A gene expression atlas of the domestic pig. *BMC Biol.* 10, 90. doi:10.1186/1741-7007-10-90.

Halstead, M. M., Kern, C., Saelao, P., Chanthavixay, G., Wang, Y., Delany, M. E., et al. (2020a). Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Sci. Rep.* 10, 5230. doi:10.1038/s41598-020-61678-9.

Halstead, M. M., Kern, C., Saelao, P., Wang, Y., Chanthavixay, G., Medrano, J. F., et al. (2020b). A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *bioRxiv*, 2020.08.13.249870. doi:10.1101/2020.08.13.249870.

Halstead, M. M., Ma, X., Zhou, C., Schultz, R. M., and Ross, P. J. (2020c). Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. *Nat. Commun.* 11, 1–16. doi:10.1038/s41467-020-18508-3.

Han, K., Ren, R., Cao, J., Zhao, S., and Yu, M. (2019). Genome-Wide Identification of Histone Modifications Involved in Placental Development in Pigs . *Front. Genet.* 10, 277.

Harrison, P. W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., et al. (2018). FAANG, establishing metadata standards, validation and best practices for the farmed

and companion animal community. *Anim. Genet.* 49, 520–526. doi:10.1111/age.12736.

Hickey, J. M., Bruce, C., Whitelaw, A., and Gorjanc, G. (2016). Promotion of alleles by genome editing in livestock breeding programmes. *J. Anim. Breed. Genet.* 133, 83–84. doi:10.1111/jbg.12206.

Jenko, J., Gorjanc, G., Cleveland, M. A., Varshney, R. K., Whitelaw, C. B. A., Woolliams, J. A., et al. (2015). Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. *Genet. Sel. Evol.* 47, 55. doi:10.1186/s12711-015-0135-3.

John M. Gaspar (2020). Genrich: detecting sites of genomic enrichment. Available at: https://github.com/jsh58/Genrich.

Johnsson, M., Gaynor, R. C., Jenko, J., Gorjanc, G., de Koning, D.-J., and Hickey, J. M. (2019). Removal of alleles by genome editing (RAGE) against deleterious load. *Genet. Sel. Evol.* 51, 14. doi:10.1186/s12711-019-0456-8.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* 9, e1003118–e1003118. doi:10.1371/journal.pcbi.1003118.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. *J. Stat. Softw.* 25, 253–258. doi:10.18637/jss.v025.i01.

Lehnert, S. A., Reverter, A., Byrne, K. A., Wang, Y., Nattrass, G. S., Hudson, N. J., et al. (2007). Gene expression studies of developing bovine longissimusmuscle from two different beef cattle breeds. *BMC Dev. Biol.* 7, 95. doi:10.1186/1471-213X-7-95.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–93. doi:10.1093/bioinformatics/btr509.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, Y., Li, B., Yang, M., Han, H., Chen, T., Wei, Q., et al. (2020). Genome-Wide Association Study and Fine Mapping Reveals Candidate Genes for Birth Weight of Yorkshire and Landrace Pigs . *Front. Genet.* 11, 183.

Lim, K. S., Lee, K. T., Park, J. E., Chung, W. H., Jang, G. W., Choi, B. H., et al. (2017). Identification of differentially expressed genes in longissimus muscle of pigs with high and low intramuscular fat content using RNA sequencing. *Anim. Genet.* 48, 166–174. doi:10.1111/age.12518.

Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., et al. (2019). An integrated chromatin

accessibility and transcriptome landscape of human pre-implantation embryos. *Nat. Commun.* 10, 364. doi:10.1038/s41467-018-08244-0.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G., et al. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv*, 085050. doi:10.1101/085050.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20. doi:10.1101/gr.107524.110.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. Genetics and population analysis BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. doi:10.1093/bioinformatics/btw044.

Oksbjerg, N., Gondret, F., and Vestergaard, M. (2004). Basic principles of muscle development and growth in meat-producing mammals as affected by the insulin-like growth factor (IGF) system. *Domest. Anim. Endocrinol.* 27, 219–240. doi:https://doi.org/10.1016/j.domaniend.2004.06.007.

Pai, A. A., Pritchard, J. K., and Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. 11, e1004857. doi:10.1371/journal.pgen.1004857.

Pardo, C. E., Kreuzer, M., and Bee, G. (2013). Effect of average litter weight in pigs on growth performance, carcass characteristics and meat quality of the offspring as depending on birth weight. *animal* 7, 1884–1892. doi:DOI: 10.1017/S1751731113001419.

Patterson, M. D., Marschall, T., Pisanti, N., Van Iersel, L., Stougie, L., Klau, G. W., et al. (2015). WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* 22, 498–509. doi:10.1089/cmb.2014.0157.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.

R Core Team (2017). R: A language and environment for statistical computing. *R Found. Stat. Comput.* Available at: https://www.r-project.org/.

Rehfeldt, C., and Kuhn, G. (2006). Consequences of birth weight for postnatal growth performance and carcass quality in pigs as related to myogenesis1. *J. Anim. Sci.* 84, E113–E123. doi:10.2527/2006.8413_supplE113x.

Rudar, M., Fiorotto, M. L., and Davis, T. A. (2019). Regulation of Muscle Growth in Early Postnatal Life in a Swine Model. *Annu. Rev. Anim. Biosci.* 7, 309–335. doi:10.1146/annurev-animal-020518-115130.

Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics* 15, 442. doi:10.1186/1471-2164-15-442.

Salavati, M., Bush, S. J. S. J., Palma-Vera, S., McCulloch, M. E. B. M. E. B., Hume, D. A. D. A., and Clark, E. L. E. L. (2019). Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression in Crossbred Sheep. *Front. Genet.* 10, 619122. doi:10.3389/fgene.2019.00863.

Soneson, C., Love, M. I., and Robinson, M. D. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research*. doi:10.12688/F1000RESEARCH.7563.2.

Stange, K., Miersch, C., Sponder, G., and Röntgen, M. (2020). Low birth weight influences the postnatal abundance and characteristics of satellite cell subpopulations in pigs. *Sci. Rep.* 10, 6149. doi:10.1038/s41598-020-62779-1.

Stenhouse, C., Hogg, C. O., and Ashworth, C. J. (2018). Associations between fetal size, sex and both proliferation and apoptosis at the porcine feto-maternal interface. *Placenta* 70, 15–24. doi:10.1016/j.placenta.2018.08.006.

Summers, K. M., Bush, S. J., Wu, C., Su, A. I., Muriuki, C., Clark, E. L., et al. (2020). Functional Annotation of the Transcriptome of the Pig, Sus scrofa, Based Upon Network Analysis of an RNAseq Transcriptional Atlas . *Front. Genet.* 10, 1355.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi:10.1093/bib/bbs017.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82. doi:10.1038/nature11232.

Tyler, T. W., and Girke, T. (2016). systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics* 17, 388. doi:10.1186/s12859-016-1241-0.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063. doi:10.1038/nmeth.3582.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43, 11.10.1-33. doi:10.1002/0471250953.bi1110s43.

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth. New York: Springer Available at: http://www.stats.ox.ac.uk/pub/MASS4/.

Wang, X., Liu, X., Deng, D., Yu, M., and Li, X. (2016). Genetic determinants of pig birth

weight variability. *BMC Genet.* 17, S15. doi:10.1186/s12863-015-0309-6.

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* 9, 1–14. doi:10.1093/gigascience/giaa051.

Weatherall, E. L., Avilkina, V., Cortes-Araya, Y., Dan-Jumbo, S., Stenhouse, C., Donadeu, F. X., et al. (2020). Differentiation Potential of Mesenchymal Stem/Stromal Cells Is Altered by Intrauterine Growth Restriction . *Front. Vet. Sci.* 7, 809.

Weber, F. E., VAUGHAN, K. T., REINACH, F. C., and FISCHMAN, D. A. (1993). Complete sequence of human fast-type and slow-type muscle myosin-binding-protein C (MyBP-C): Differential expression, conserved domain structure and chromosome assignment. *Eur. J. Biochem.* 216, 661–669. doi:10.1111/j.1432-1033.1993.tb18186.x.

Wigmore, P. M., and Stickland, N. C. (1983). Muscle development in large and small pig fetuses. *J. Anat.* 137, 235–45.

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., et al. (2012). Tissue sampling methods and standards for vertebrate genomics. *Gigascience* 1, 8. doi:10.1186/2047-217X-1-8.

Wu, J., Xu, J., Liu, B., Yao, G., Wang, P., Lin, Z., et al. (2018). Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* 557. doi:10.1038/s41586-018-0080-8.

Xiang, R., Berg, I. van den, MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci.* 116, 19398 LP – 19408. doi:10.1073/pnas.1904159116.

Xue, H.-L. L., and Zhou, Z.-X. X. (2006). Effects of the MyoG Gene on the Partial Growth Traits in Pigs. *Acta Genet. Sin.* 33, 992–997. doi:10.1016/S0379-4172(06)60134-0.

Yan, L., Guo, H., Hu, B., Li, R., Yong, J., Zhao, Y., et al. (2016). Epigenomic Landscape of Human Fetal Brain, Heart, and Liver. *J. Biol. Chem.* 291, 4386–4398. doi:10.1074/jbc.M115.672931.

Yang, X. R., Yu, B., Mao, X. B., Zheng, P., He, J., Yu, J., et al. (2015). Lean and obese pig breeds exhibit differences in prenatal gene expression profiles of muscle development. *Animal* 9, 28–34. doi:DOI: 10.1017/S1751731114002316.

Zhao, X., Mo, D., Li, A., Gong, W., Xiao, S., Zhang, Y., et al. (2011). Comparative Analyses by Sequencing of Transcriptomes during Skeletal Muscle Development between Pig Breeds Differing in Muscle Growth Rate and Fatness. *PLoS One* 6, e19774.

Zhao, Y., Li, J., Liu, H., Xi, Y., Xue, M., Liu, W., et al. (2015). Dynamic transcriptome profiles of skeletal muscle tissue across 11 developmental stages for both Tongcheng and Yorkshire pigs. *BMC Genomics* 16, 377. doi:10.1186/s12864-015-1580-7.

Figure Legends

Figure 1: Flowchart describing the experimental design and samples included in each stage of the analysis performed in this study. Colour coding indicates where there are overlaps in the analysis performed for each component of the study.

Figure 2: Location of ATAC-Seq peaks within genomic features (A) and distance of ATAC-Seq peaks from the TSS (B) in all 24 samples included in the study. The samples are sorted by the developmental timeline (Day 45 to 6 weeks old from top to bottom) in both panels. The cryopreserved nuclei samples are only present in the 6 week old group and are labelled as cryo_nuclei.

Figure 3: The width distribution of ATAC-Seq peaks at each developmental time point. The dotted line represents the median width for each time point. A) Width of ATAC-Seq peaks shared across all time points, B) Width of ATAC-Seq peaks that were specific to each time point. WO stands for week old.

Figure 4: Mapping of ATAC-Seq peaks relative to the TSS, averaged across the three sizes of piglet, showing variation in peak width across the five developmental time points analysed.

Figure 5: Mapping of ATAC-Seq peaks at Day 90 of gestation that were: i) specific to small or large sized piglets or ii) shared across samples. A) mapping relative to the TSS, B) distribution within genomic features and C) distribution of relative to the TSS.

Figure 6: Principal Component Analysis (PCA) of gene expression estimates (as TPM) from the RNA-Seq data for each sample. The samples cluster according to developmental stage with clear separation of neonatal and post-natal samples. D45 = Gestational Day 45; D60 = Gestational Day 60; D90 = Gestational Day 90; 1WK = Neonatal 1 week old ;6WK = Juvenile 6 weeks old.

Figure 7: Differentially expressed genes between large, average and small sized piglets at Day 90 of gestation. A) Large vs Small B) Average vs Small C) Average vs Large. Differentially expressed genes are shown in red, the log2FC >0.1 in blue and the significance threshold as a green line. The ApeGLM shrinkage method was used to normalised the log fold change plotted on the x axis. Detailed methodology is described in (Love et al., 2014).

Figure 8: Proximity of ATAC-Seq peaks specific to large (A) and small (B) piglets and differentially expressed genes. Differentially expressed genes are marked in green. The x-axis (start_dist) is the distance from the start of the gene model to the start of the ATAC-Seq peak, for +ve values the peak is either within the gene or within 10kbp of the 3’ end and for -ve values the peak is within 10Kbp of the 5’ end of the gene. The y-axis indicates the width of the peak. As the y axis represent the width of the peak, the larger the node the wider the ATAC-Seq peak.

Figure 9: The overlay of DNA openness score (DOS) and transcript expression level (as TPM) in large and small sized piglets from three litters: A & B) Small and large piglets from litter 251016. C & D) small and large piglets from litter 111016. E) small piglet from 231115 (no RNA-Seq sample from a large piglet was available for this litter from comparison) and F) large piglet from 121016 (no RNA-Seq sample from a small piglet was available for this litter from comparison). Dots are coloured and sized according to increasing TPM. Colours represent DOS binned (10 bins) categories of 0.8-1 (purple), 0.5-0.8 (blue), 0.2-0.5 (green) and 0-0.2 in yellow. The average (mean) expression of transcripts in each sample marked by a red dotted line.

Figure 10: Genes exhibiting significant allele specific expression in small, average and large sized piglets at Day 90 of gestation. A) genes exhibiting allelic imbalance in up to three biological replicates for each size of piglet (genes were classified into three groups according to whether allelic balance was detected in all three biological replicates: 3= exhibited in large, average and small, 2= exhibited in either small and large or average and large, 1= exhibited only in one group). The percentage representation (75% to 100 %) of the gene exhibiting allelic imbalance across biological replicates is indicated by the transparency of the bars. The genes showing ASE in each of the three sizes of piglet are indicated using coloured bars that represent the level of allelic imbalance calculated using the Liptak score (and output metric of the GeneiASE pipeline). B) Distance of ATAC-Seq peaks, specific to either large or small sized piglets, from the TSS of genes showing significant allelic imbalance.

Figure 11: Genomic track visualisation of the genomic regions including gene *ARRDC3*. Variants called using RNA-Seq data with read-backed haplotype blocks in the small sized piglets, and the corresponding open chromatin regions (ATAC-Seq peaks), shown. The top four tracks show variants in dark blue (one consensus genotype and three allele-specific positive variants in each piglet) and the haplotype block in green. ATAC-Seq peaks specific to small sized piglets are shown in pink and shared ATAC-Seq peaks in black.

Figure 12: Genomic track visualisation of the genomic region including gene *PXDN.* Variants called using RNA-Seq data with read-backed haplotype blocks in the small sized piglets, and the corresponding open chromatin regions (ATAC-Seq peaks), are shown. The top four tracks show variants in dark blue (one consensus genotype track and three separate tracks for allele-specific positive variants in each of the three piglets) and the haplotype block in green. ATAC-Seq peaks specific to small sized piglets are shown in pink and shared ATAC-Seq peaks in black.

Supplemental Material

Supplementary Table 1: Weights and crown-rump lengths of all piglets included in the study.

Supplementary Table 2: Details of components of all buffers used for ATAC-Seq sample collection and library preparation.

Supplementary Table 3: Details of the primers (Ad2.x variable index) used for generating each ATAC-Seq library. Each sample is barcoded with a different variable index.

Supplementary Table 4: Summary table of RNA-sequencing and ATAC-sequencing quality control metrics for all 28 samples. 'N/A' indicates that sequencing did not occur for that sample.

Supplementary Table 5: Results of differential expression analysis of large vs average vs small piglet sizes at Day 90 of gestation using a DESeq2 glm model. Only genes with FDR < 0.1 have been included.

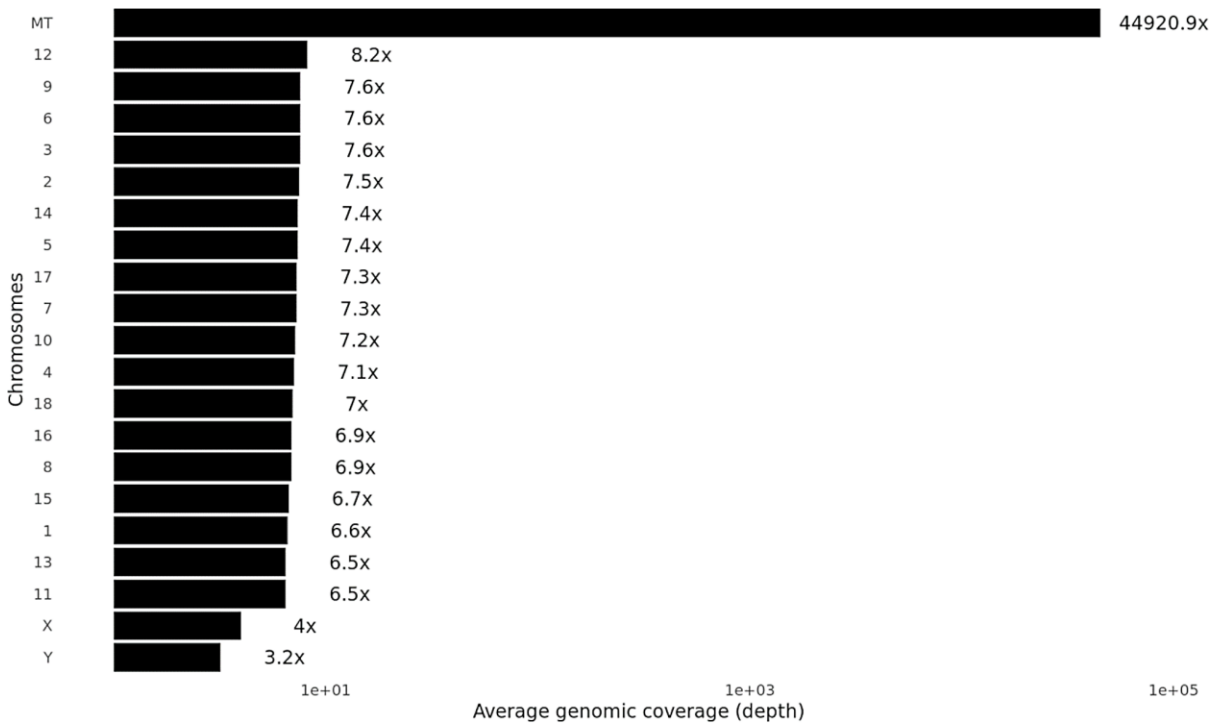
Supplementary_file_1.zip: The collection of scripts and BED files corresponding to the ATAC-Seq peak coordinates for the developmental time points and sizes of piglet at Day 90.

Supplementary_file_2.zip: Library preparation and sequencing metrics for all ATAC-Seq samples generated for this study.

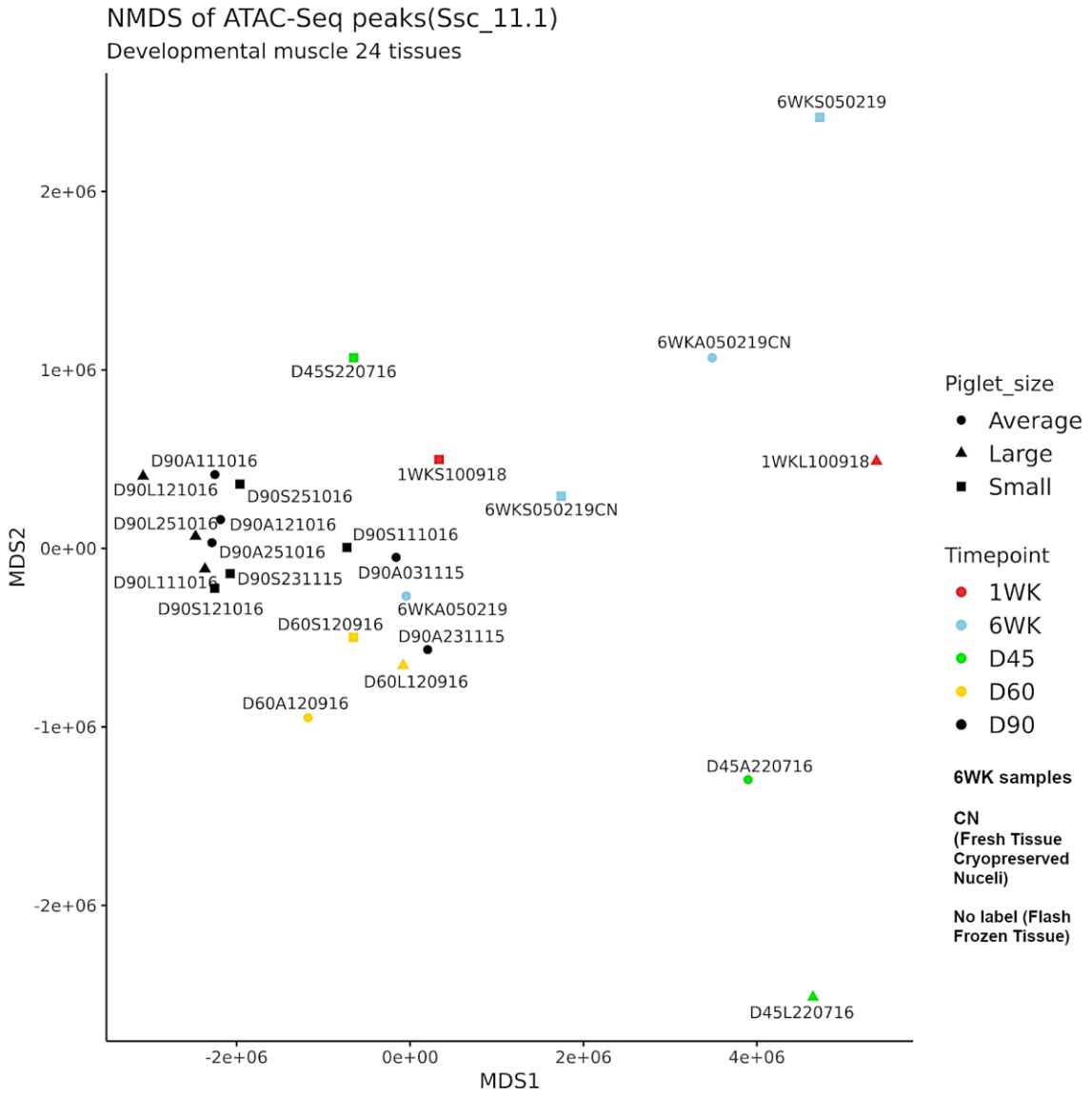
Supplementary_file_3.zip: The haplotype specific ORF predictions and corresponding genes in small piglets. Modified BED format presented as a tab separated output.

Supplementary_file_4.zip: The library preparation metadata and sequencing read depth metrics of the ATAC-Seq samples (n=24).

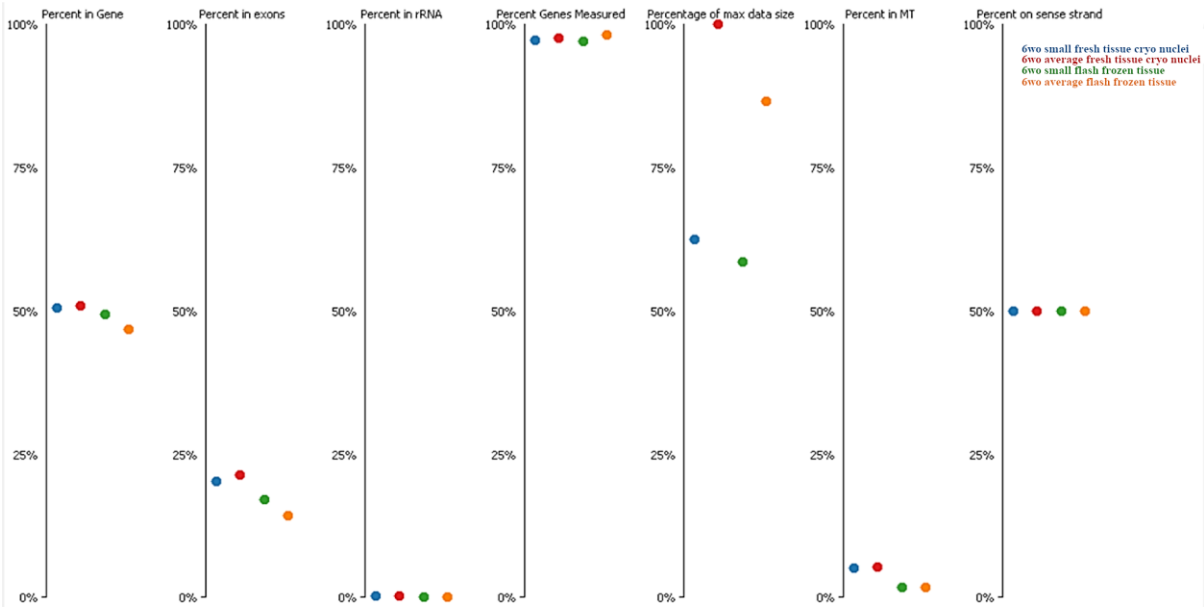
Supplementary_file_5.zip: GTF track of Day 90 small piglet haplotype blocks phased using an RNA-Seq read-back phasing algorithm in the WhatsHap pipeline with fasta files for each individual piglet.



Supplementary Figure S1: Average genomic coverage of the ATAC-Seq dataset by chromosome.

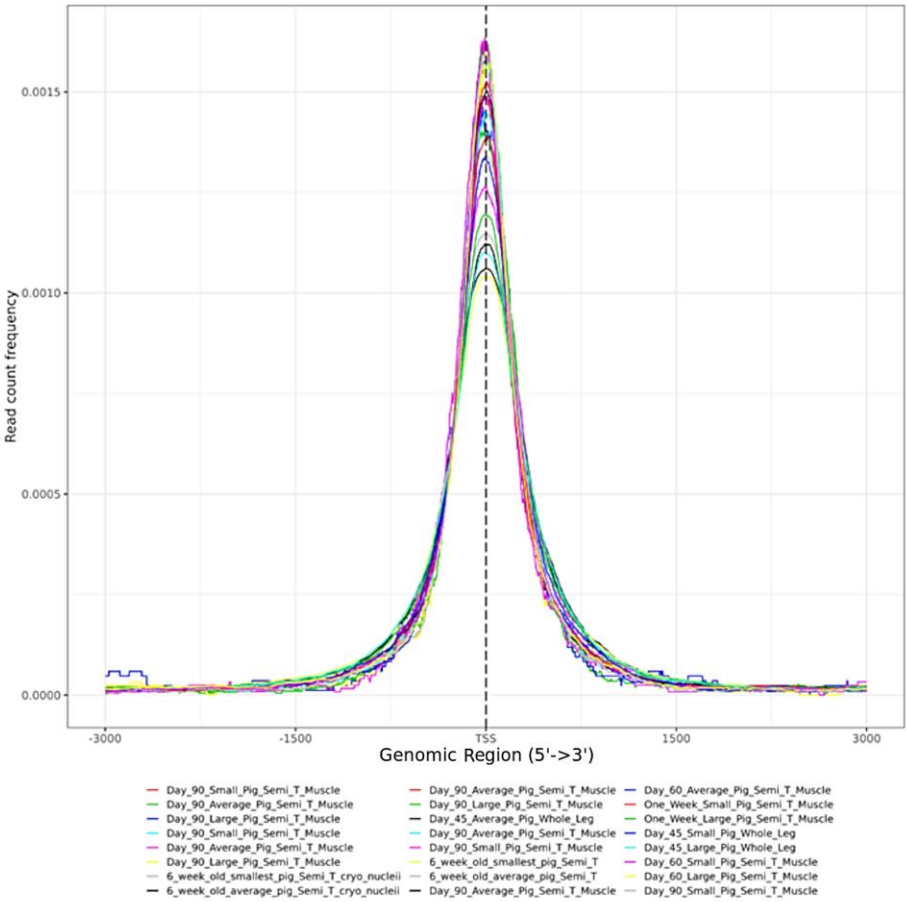


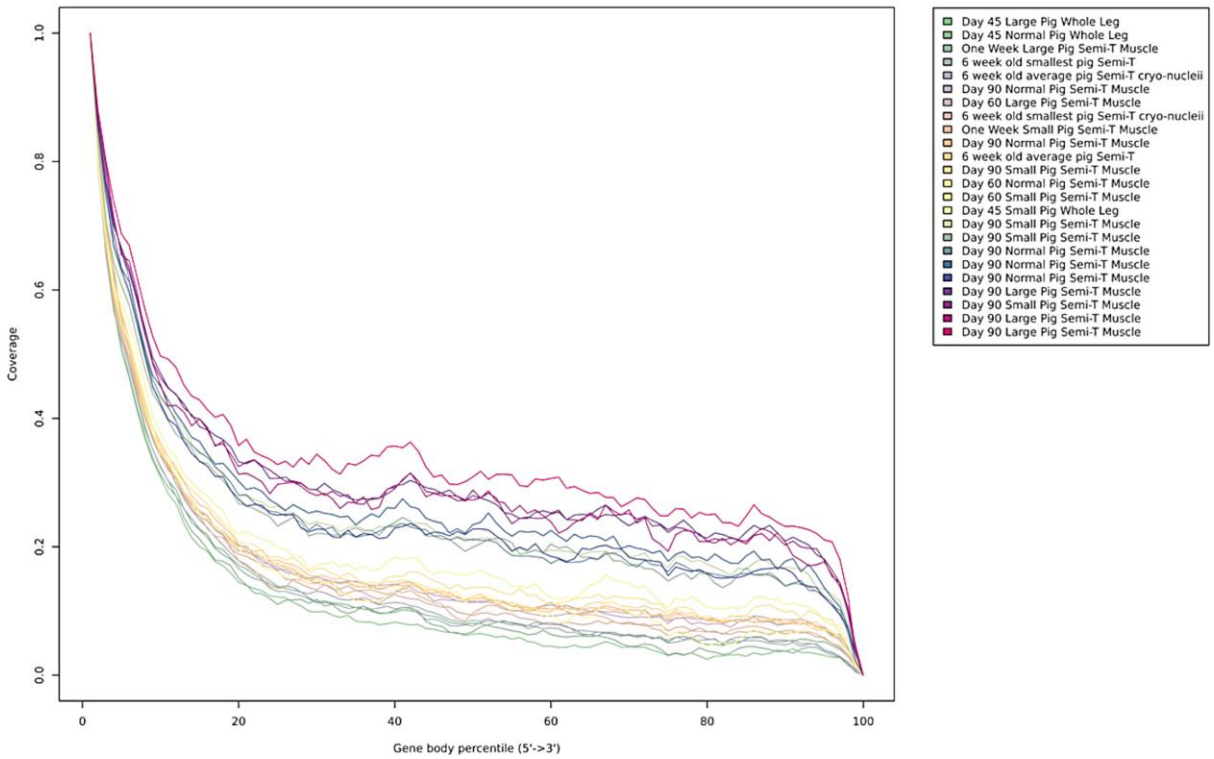
Supplementary Figure S2: Non-linear Multi-dimensional scaling (NMDS) analysis of the ATAC-Seq open chromatin regions annotated at transcript level in all samples. The signal (AUC value) for each transcript was calculated based on the peaks present in ±3 Kbp vicinity of its promoter region. Samples from different developmental time points are indicated by colour and piglet size by shape. CN stands for cryopreserved nuclei samples in the 6 week-old samples.



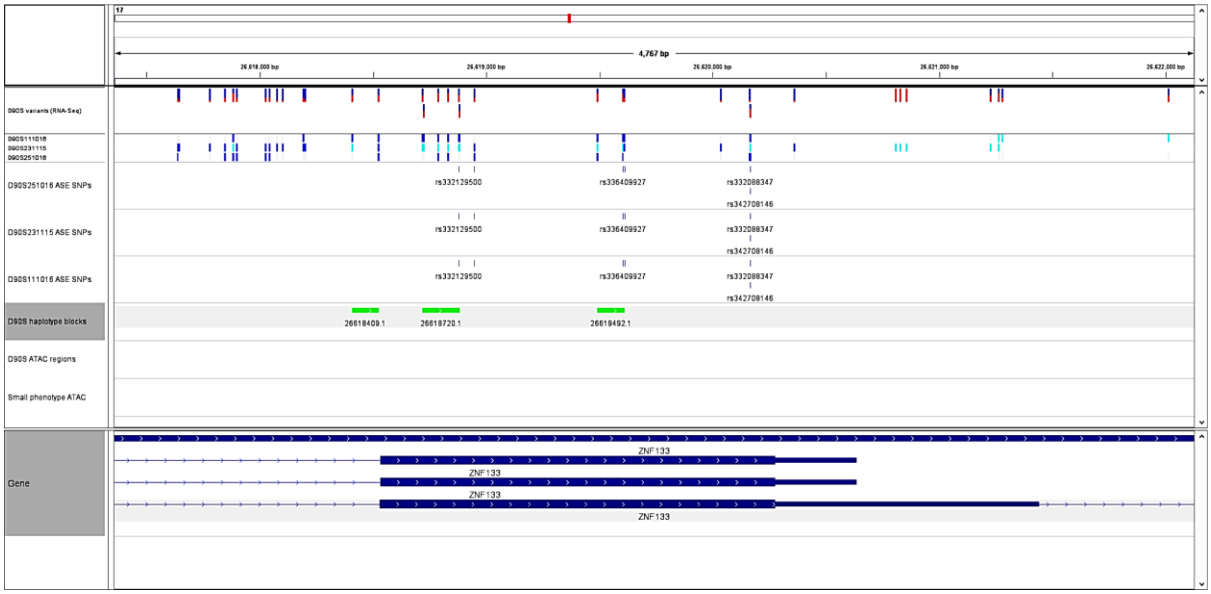
Independent two sided t-test (CN vs fresh tissue)	p value
Percent in Gene	0.2989475
Percent in exons	0.1221280
Percent in rRNA	0.1052326
Percent Genes Measured	0.8628498
Percent of max data size	0.7510646
Percent in MT	**0.0146693**
Percent on sense strand	0.9947984

Supplementary Figure S3: Metrics to compare ATAC-Seq peaks in libraries prepared from either cryopreserved nuclei from fresh tissue (red and blue) or from flash frozen tissue (orange and green). Both sets of samples were collected from the semitendinosus muscle from the same six-week-old piglets. Table shows the result of a t-test performed between the two library preparation methods in each category. MT read alignments from the libraries prepared from cryopreserved nuclei (5.2%) were significantly higher than fresh tissue (1.7%).





Supplementary Figure S4: Open chromatin peak distribution from the ATAC-Seq libraries generated for all 24 tissue samples. The read distribution was calculated based on the distance from the centre of the ATAC-Seq peak (highest read count position) to TSSs.



Supplementary Figure S5: Genomic track visualisation of the genomic region including gene *ZNF133.* Variants called using RNA-Seq data with read-backed haplotype blocks in the small sized piglets, and the corresponding open chromatin regions (ATAC-Seq peaks), are shown. The top four tracks show variants in dark blue (one consensus genotype and three ASE positive variants in each piglet) and the haplotype block in green. ATAC-Seq peaks specific to small sized piglets are shown in pink and shared ATAC-Seq peaks in black.