DISS. ETH NO. 29579

IDENTIFICATION OF GENETIC CHARACTERISTICS AND TRAIT-ASSOCIATED VARIANTS IN SWISS PIGS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES

(Dr. sc. ETH Zurich)

presented by

ADÉLA POUBLAN-COUZARDOT

Ing., Czech University of Life Sciences Prague

born on 28.10.1993

citizen of the Czech Republic

accepted on the recommendation of

Prof. Dr. Hubert Pausch

Prof. Dr. Christine Baes

2023

Table of Contents

Acknowledgements

I would like to express my sincere gratitude and appreciation to everyone who has contributed to the successful completion of this dissertation.

To my supervisor Prof. Dr. Hubert Pausch, whose unwavering guidance, encouragement, and invaluable insights have been instrumental throughout this research journey. I appreciate your feedback, even if deciphering your comments sometimes felt like cracking secret codes. I promise I didn't take your critiques personally... well, most of the time.

To PD Dr. Stefan Neuenschwander, my co-supervisor, who was always prepared to help, and whose funding made it all possible.

To Prof. Dr. Christine Baes who joined my thesis committee and reviewed my thesis.

To SUISAG, with special thanks to Dr. Andreas Hofer and Dr. Negar Khayatzadeh, who shared valuable insights about the pig breeding in Switzerland, Micarna SA and the ETH Zürich Foundation for funding.

To the strangers who became friends and the friends who became family during the past four years, your support and encouragement have made the research environment more vibrant, enjoyable, and have warmed the coldness of a foreign country. Special thanks to Dr. Alexander Leonard, Audald Lloret i Villas, Dr. Naveen Kumar Kadri, and Xena Marie Mapel for proofreading this thesis and providing extra help whenever I needed it.

To Iveta and Miroslav, my parents, thank you for your unwavering support, understanding, and motivation.

To Pierre, my husband, thank you for standing by me through the highs and lows of this pursuit, reminding me of what truly matters.

To Gabriel, my son, for teaching me efficiency and speed writing during his (thankfully frequent) naps.

Thank you all.

Adéla

Summary

A typical Swiss fattening pig inherits 70% of its genes from animals of the dam and sire line of the Swiss Large White (SLW) breed, i.e., two domestic lines that are bred in Switzerland to meet the pork market requirements and societal expectations. Although knowledge on breed-specific sequence variants is essential for genomic breeding and detection of deleterious alleles, the Swiss pig populations have not yet been thoroughly genetically characterized.

In Chapter 2, key ancestors from two SLW lines were sequenced using short reads to assess genetic differentiation between the breeds. Sequencing the genomes of 70 boars with an average coverage of 16.69-fold allowed for the assessment of sequence variation in the form of single nucleotide and small insertion and deletion polymorphisms. Principal component, admixture, and fixation index analyses indicated significant genetic differentiation between the lines. Genomic inbreeding, quantified through runs of homozygosity, was found to be higher in the sire line compared to the dam line. Additionally, 51 signatures of selection were detected using two complementary approaches, although only six overlapped between the lines. By using the sequenced haplotypes of the 70 key ancestors as a reference panel, genotypes were called in 175 low-coverage sequences using the GLIMPSE software. The genotype concordance of 97.60%, non-reference sensitivity of 98.73%, non-reference discrepancy of 3.24% between the inferred genotypes and those obtained from Illumina PorcineSNP60 BeadChip demonstrated high

accuracy. These findings shed light on the genetic diversity within and between the SLW pig lines, providing insights into the potential of low-pass sequencing.

Chapter 3 explored genetic correlations between complex traits and applied genome-wide association studies (GWAS) to detect trait-associated markers. By considering array-derived genotypes and phenotypes for 24 reproduction, production, and conformation traits, the study aimed to compare the outcomes of multivariate association testing and multi-trait meta-analysis with single-trait GWAS. A cohort of 5,753 pigs from the SLW breed was examined, analyzing genotypes at 44,733 SNPs. Single-trait association analyses identified eleven quantitative trait loci (QTL) influencing 15 traits. The two multi-trait methods revealed between 3 and 6 QTL within each of the three trait groups. The results demonstrated that both methods produce similar outcomes, though each had unique advantages. The investigation of single-trait association studies proved valuable in identifying specific QTL, while multi-trait methods provided an overview of pleiotropic loci. Array-derived genotypes were imputed to sequence level using a reference panel of 421 pigs, yielding imputed genotypes for 16 million sequence variants with high accuracy. Fine-mapping of six QTL with the imputed sequence variant genotypes highlighted four previously proposed causal mutations among the top variants.

In Chapter 4, the branch point sequence, a degenerate intronic heptamer critical for spliceosome assembly during pre-mRNA splicing, was examined. Despite its functional importance, the branch point

sequence is often omitted from genome annotations, including the pig annotation. The study aimed to predict branch point sequences in 30 plant and animal species, including pigs, and assess their evolutionary constraints using public variant databases. Our analysis revealed an irregular distribution of variants in public databases that affected 16 out of 30 species investigated, predominantly due to biased or erroneous variants. In 14 species with largely unbiased databases, from which three (pig, goat and sheep) had in-house whole-genome sequence variants, evolutionary constraint analysis demonstrated that the fourth and sixth positions of the branch point sequence were stronger constrained than coding nucleotides. The study highlighted the need to scrutinize public variant databases for potential biases before relying on them for genomic analyses.

Overall, this doctoral thesis provides insights into genetic analysis, trait variation, and evolutionary constraints within the SLW pig breed. Specifically, this research highlights the necessity of understanding genetic diversity, the benefits of employing different analytical approaches and variation sources, and the importance of accurate annotation and analysis of genetic variants in evolutionary studies.

Zusammenfassung

Ein typisches Schweizer Mastschwein erbt 70% seiner Gene von Tieren der Mutter- und Vaterlinie der Rasse Swiss Large White (SLW), d. h. von zwei heimischen Linien, die in der Schweiz gezüchtet werden, um die Anforderungen der Konsumenten und die gesellschaftlichen Erwartungen zu erfüllen. Obwohl die Kenntnis rassespezifischer Sequenzvarianten für die genomische Zucht und die Erkennung schädlicher Allele wichtig ist, sind die Schweizer Schweinepopulationen noch nicht umfassend genetisch charakterisiert.

In Kapitel 2 wurden die wichtigsten Vorfahren aus den beiden Linien der Rasse Swiss Large White mit kurzen DNA Fragmenten genomweit sequenziert. Die Sequenzierung der Genome von 70 Ebern mit einer durchschnittlichen Abdeckung von 16,69-fach ermöglichte es, Sequenzvariation in Form von Einzelnukleotid- und kleinen Insertions- und Deletionspolymorphismen zu charakterisieren. Verschiedenen populationsgenetische Analysen zeigten eine signifikante genetische Differenzierung zwischen den Linien. Die genomische Inzucht, quantifiziert durch sogenannte «runs of homozygosity», war in der Vaterlinie höher als in der Mutterlinie. Darüber hinaus wurden mit zwei komplementären Ansätzen 51 Selektionssignaturen festgestellt, von denen allerdings nur sechs zwischen den Linien überlappten. Unter Verwendung der sequenzierten Haplotypen der 70 wichtigsten Vorfahren als Referenzpanel wurden mit der GLIMPSE-Software Genotypen in 175 Sequenzen mit geringer Abdeckung bestimmt. Eine Übereinstimmung von 97,60 %, die Nicht-Referenz-Sensitivität von 98,73 % und die Nicht-

Referenz-Diskrepanz von 3,24 % zwischen den abgeleiteten Genotypen und den aus dem Illumina PorcineSNP60 BeadChip gewonnenen Genotypen belegen eine hohe Genauigkeit dieses Vorgehens. Diese Ergebnisse geben Aufschluss über die genetische Vielfalt innerhalb und zwischen den Rassen der Schweizer Large White Schweine und liefern Einblicke in das Potenzial der Low-Pass-Sequenzierung.

Kapitel 3 untersuchte genetische Korrelationen zwischen komplexen Merkmalen und berichtet über genomweiter Assoziationsstudien (GWAS) zur Erkennung von Merkmals-assoziierten Markern. Unter Berücksichtigung der von Arrays abgeleiteten Genotypen und Phänotypen für 24 Reproduktions-, Produktions- und Exterieur-Merkmale, werden die Ergebnisse von multivariaten Assoziationstests und Meta-Analysen für mehrere Merkmale mit GWAS für einzelne Merkmale zu vergleichen. Eine Kohorte von 5.753 Schweinen der Rasse Swiss Large White wurde untersucht, wobei Genotypen an 44.733 SNPs analysiert wurden. Die Assoziationsanalysen einzelner Merkmale identifizierten elf quantitative Merkmalsloci (QTL), die 15 Merkmale beeinflussen. Die beiden Mehrmerkmals-Methoden spürten in den drei Merkmalsgruppen zwischen 3 und 6 QTL auf. Die Ergebnisse zeigten, dass beide Methoden zwar ähnliche Ergebnisse liefern, aber jede ihre Vorteile hat. Die Untersuchung von Single-Trait-Assoziationsstudien erwies sich als wertvoll für die Identifizierung spezifischer quantitativer Merkmalsloci (QTL), während Multi-Trait-Methoden einen Überblick über pleiotrope Loci lieferten. Die von den Arrays abgeleiteten Genotypen wurden mit Hilfe eines Referenzpanels von 421 Schweinen auf Sequenzebene imputiert, was zu imputierten Genotypen für 16 Millionen Sequenzvarianten mit hoher

Genauigkeit führte. Die Feinkartierung von sechs QTL unter Verwendung von imputierten Sequenzgenotypen spürte vier zuvor vorgeschlagene kausale Mutationen unter den Top-Varianten auf.

In Kapitel 4 wurde die Branch-Point-Sequenz, ein degeneriertes intronisches Heptamer, das für den Zusammenbau des Spleißosoms während des Spleißens der prä-mRNA entscheidend ist, untersucht. Trotz ihrer funktionellen Bedeutung wird dieses regulatorische Element meist vernachlässigt, so auch in der Annotation des Schweinegenoms. Ziel der Studie war die Vorhersage der Branch-Point-Sequenz in 30 Pflanzen- und Tierarten, einschließlich Schweinen, und die Bewertung ihrer Variabilität anhand öffentlich zugängiger Variantendatenbanken. Unsere Analyse ergab, dass 16 der 30 untersuchten Variantendatenbanken verzerrt oder unvollständig waren, was in erster Linie auf verzerrte oder fehlerhafte Varianten zurückzuführen ist. In 14 Arten mit weitgehend unverzerrten Datenbanken, von denen für drei (Schwein, Ziege und Schaf) auch genomweite Sequenzdaten zur Verfügung standen, waren sowohl die vierte wie auch sechste Position der Branch-Point-Sequenz weniger variabel als kodierende Nukleotide. Die Studie unterstreicht die Notwendigkeit, öffentliche Variantendatenbanken auf mögliche Verzerrungen hin zu überprüfen, bevor man sich bei genomischen Analysen auf sie verlässt.

Insgesamt bietet diese Doktorarbeit Einblicke in die genetische Architektur, die Merkmalsvariation und die evolutionären Zwänge innerhalb einer Schweizer Schweinepopulation. Die Forschung unterstreicht die Bedeutung des Verständnisses der genetischen Vielfalt, die Vorteile des Einsatzes verschiedener analytischer Ansätze und

Variationsquellen sowie die Bedeutung einer genauen Annotation und Analyse genetischer Varianten in der Evolution.

Résumé

Le patrimoine génétique d'un porc d'engraissement suisse typique provient pour 70% des lignées maternelle et paternelle de la race Swiss Large White (SLW), c'est-à-dire de deux lignées domestiques élevées en Suisse pour répondre aux exigences du marché de la viande porcine et aux attentes de la société. Bien que la connaissance des variants de séquence spécifiques à la race soit essentielle pour la sélection génomique et la détection des allèles délétères, les populations porcines suisses n'ont pas encore été caractérisées génétiquement de manière approfondie.

Dans le Chapitre 2, des ancêtres clés des deux lignées de la race SLW ont été séquencés à l'aide de « courtes lectures ». Le séquençage des génomes de 70 verrats avec une couverture moyenne de 16,69X a permis d'évaluer les variants génétiques sous forme de polymorphismes nucléotidiques (SNP - *Single Nucleotide Polymorphism* -) et de petits polymorphismes d'insertion et de délétion. L'analyse en composantes principales, l'analyse des mélanges et l'indice de fixation ont indiqué une différenciation génétique significative entre les lignées. La consanguinité génomique, quantifiée par les profils d'homozygotie, s'est avérée plus élevée dans la lignée des pères que dans celle des mères. En outre, 51 signatures de sélection ont été détectées à l'aide de deux approches complémentaires, bien que seules 6 d'entre elles se chevauchaient entre les lignées. En utilisant les haplotypes séquencés des 70 ancêtres clés comme panel de référence, les génotypes ont été appelés dans 175 séquences à faible couverture à l'aide du logiciel GLIMPSE. La concordance génotypique de 97,60 %, la sensibilité non référentielle de 98,73 %, et la

divergence non référentielle de 3,24 % entre les génotypes inférés et ceux obtenus à partir de la puce Illumina PorcineSNP60 BeadChip, ont démontré une grande précision de l'imputation. Ces résultats mettent en lumière la diversité génétique au sein de la race porcine SLW et entre les deux lignées, et donnent un aperçu du potentiel du séquençage à faible couverture.

Le Chapitre 3 explore les corrélations génétiques entre caractères complexes et l'application des études d'association pangénomiques (GWAS - *Genome-Wide Association Study* -) pour détecter les marqueurs associés aux caractères. À l'aide de génotypes dérivés des puces et de phénotypes de 24 caractères de reproduction, de production et de conformation, l'étude visait à comparer les résultats des tests d'association multivariés et des méta-analyses multi-caractères avec ceux des GWAS à caractère unique. Une cohorte de 5 753 porcs de la race SLW, génotypée sur 44 733 SNP, a été analysée. Les analyses d'association à un seul caractère ont permis d'identifier 11 *loci* de caractères quantitatifs (QTL - *Quantitative Trait Locus* -) influençant 15 caractères. Les deux méthodes multi-caractères ont révélé dans les trois groupes de caractères entre 3 et 6 QTL. Les résultats ont montré que les deux approches donnaient des résultats similaires mais présentaient des avantages uniques : les études d'association à un seul caractère se sont avérées utiles pour identifier des QTL spécifiques, tandis que les méthodes à plusieurs caractères ont permis d'obtenir une vue d'ensemble des *loci* pléiotropiques. Les génotypes dérivés des matrices ont été imputés au niveau de la séquence en utilisant un panel de référence de 421 porcs, permettant d'obtenir des génotypes imputés pour 16 millions de variants de séquence avec une grande précision. La cartographie fine de 6 QTL utilisant des génotypes de variants de séquence

imputés a mis en évidence 4 mutations causales précédemment proposées parmi les variants les plus importants.

Le Chapitre 4 examine la séquence du point de branchement, un heptamère intronique dégénéré essentiel à l'assemblage du splicéosome pendant l'épissage des ARN pré-messagers. Malgré son importance fonctionnelle, la séquence du point de branchement est souvent omise des annotations du génome, comme c'est le cas chez le porc. L'étude visait à prédire les séquences du point de branchement chez 30 espèces végétales et animales, dont le porc, et à évaluer leurs contraintes évolutives à l'aide de bases de données publiques de variants. Notre analyse a révélé une distribution irrégulière des variants dans les bases de données publiques pour 16 des 30 espèces étudiées, principalement en raison de variants biaisés ou erronés. Chez 14 espèces dont les bases de données sont largement exemptes de biais, et dont 3 (porc, chèvre et mouton) disposaient également de variants internes de séquence du génome entier appelés à partir d'alignements de lectures de séquences, l'analyse des contraintes évolutives a démontré que les 4ème et 6ème positions de la séquence du point de branchement étaient soumises à des contraintes plus fortes que les nucléotides codants. L'étude a mis en évidence la nécessité d'examiner minutieusement les bases de données de variants publiques pour y déceler des biais potentiels avant de s'y fier pour des analyses génomiques.

En conclusion, cette thèse de doctorat donne un aperçu de la diversité génétique, de la variation des caractères et des contraintes évolutives au sein de la race porcine SLW. Elle met en évidence l'importance de la compréhension de la diversité génétique, les avantages de l'utilisation de

différentes approches analytiques et sources de variation, et l'importance de l'annotation et de l'analyse précises des variants génétiques dans l'évolution de la race porcine.

Thesis Outline

The thesis is structured as follows:

Chapter 1 provides a literature review to introduce the Swiss pig breeds, ways of studying genomic diversity, and applications of genotypic data.

Chapter 2 characterizes the genomic diversity within and between two Swiss pig lines. This chapter has been published in *BMC Genomics*:

https://doi.org/10.1186/s12864-021-07610-5

Chapter 3 reports on associations between genotype markers and economically relevant traits relevant to the dam line and evaluates different statistical frameworks underlying the association analyses. This chapter has been published in *BMC Genomics*:

https://doi.org/10.1186/s12864-023-09295-4

Chapter 4 reports on missing feature in current annotation of genomes of different species, including *Sus scrofa*. This chapter has been deposited at the Biorxiv:

https://doi.org/10.1101/2023.03.27.534366

Chapter 5 provides a general discussion and outlook for future research.

1 General Introduction

1.1 Reference genome

Reference genomes are the cornerstone of modern genomics [1]. In farmed animal species — such as the domestic pig (*Sus scrofa*) — genome sequences have been crucial to the discovery of molecular genetic variants and the development of single nucleotide polymorphism (SNP) chips [2], which have enabled efforts to dissect the genetic underpinnings of complex traits [3]. Livestock and companion animals were among the first species to have reference genomes assembled and published: chicken (2004), dog (2004), cattle (2009), rabbit (2009), horse (2009), sheep (2010), pig (2012), and goat (2013) [4]. These high-quality genomes differed from draft genomes with their completeness (low number of gaps), low number of errors, and high percentage of sequence assembled into chromosomes.

A previously published draft pig reference genome sequence (Sscrofa10.2), developed by the Swine Genome Sequencing Consortium [5], had several deficiencies. For example, ~10% of the pig genome was not represented or incompletely represented in the assembled scaffolds, including some important genes (e.g., *CD163* and *IGF2*) [6]. The Sscrofa10.2 assembly had a scaffold N50 (i.e., the length cutoff for the longest scaffolds that contain 50% of the total genome length) value of 0.5 million bp across 9,906 scaffolds. The first major revision after the completion of the reference genome was published in 2019 as Sscrofa11.1. This current reference genome significantly improved the continuity to a scaffold N50 of 88 million bp across 706 scaffolds. Equally, the annotation

of the genome improved by doubling the number of annotated gene transcripts from 30 to 63 thousand [7].

1.2 Genomic markers and technologies

DNA technology has been applied to commercial pig breeding since the 1990s [8] in the form of a marker test. Since then, developments in sequencing technology have driven new methodologies and applications in basic and applied science. Microarray-derived genotypes are now used for improving accuracy of genomic selection [9], association testing, and fine-mapping of quantitative trait loci (QTL), which consequentially provides genomic information for phenotypes that would be cost prohibitive to measure in industry herds [10]. Examples of this would be genetic evaluation of detailed meat quality traits on production animals [11], anestrus behavior in gilts or postpartum sows, ovulation rate in sows [12], or viral disease challenges [13]. SNP chips can be also used to produce and/or validate new reference resources, for instance, in constructing a new high-density genetic linkage map [14] or assessing the completeness of the new reference sequence [7].

1.2.1 Arrays - low/high density

With the introduction of next generation sequencing platforms by Sanger in 1975 [15], high-throughput SNP discovery began in livestock species [16, 17] to create panels suitable for genotyping large numbers of DNA markers on a commercial scale [18]. The first commercial genome-wide SNP array (Illumina BovineSNP50 BeadChip) became available for cattle in 2008 [17]. Since then, a series of medium and high-density SNP chips have been developed for other livestock species such as pig, sheep,

horse, goat, and buffalo (Figure 1.1). These comprehensive panels generally contained SNP markers that are equally distributed across the genome with neutral (> 5 %) minor allele frequencies (MAF). The SNP content of these arrays is further adapted as information on individual SNP performance, thereby improving genome assemblies. The number of SNPs on one array can vary between 10,000 (e.g., low-density chip) and 777,000 (e.g., Bovine High Density [HD] chip).

Currently, the most common porcine SNP panel interrogates genotypes for 64,232 markers, with most loci lacking a known biological function [19]. Parameters, such as estimated MAF, or spacing and number of the SNPs on each chromosome, were considered during the array development. The chip contains genome-wide loci with a range of MAFs, though proprietary SNPs (i.e., any SNP linked to a patent) were excluded. The DNA samples used to select SNPs on the chip were obtained from the Duroc, Piétrain, Landrace, and Large White commercial breeds from Europe and North America, and wild boar from Japan and Europe. The

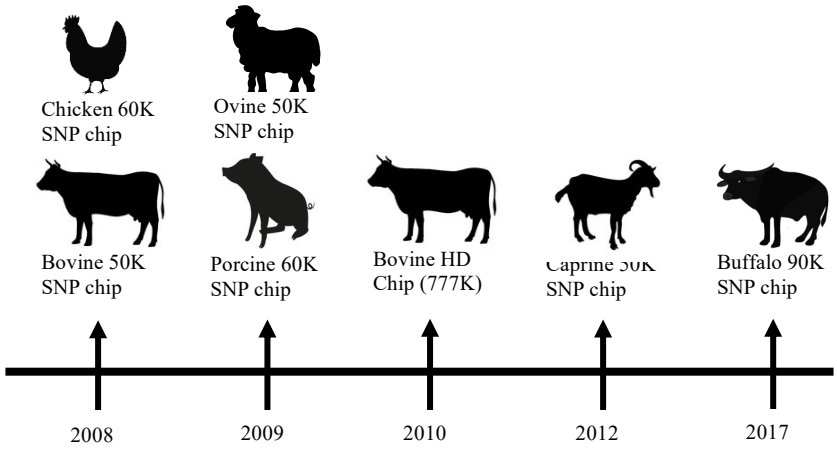


Figure 1.1: First publicly available microarrays of livestock.

resulting PorcineSNP60 Genotyping BeadChip was released at the end of 2008 [2].

1.2.2 Sequencing technologies

With the development of sequencing technologies and subsequent reductions in costs, use of whole-genome resequencing emerged as an alternative to SNP chips [20]. There are important advantages in having full-sequence compared to array-derived genotypes, including: removing SNP ascertainment bias [21], uncovering all extant variability [22–24], recovering the full unbiased demographic history of populations [22, 25], etc. Advancements in next-generation sequencing technology allowed the sequencing of larger cohorts and enabled the development of more reliable and cost-effective methods for identifying SNP variants [26] and QTLs associated with traits of economic importance (e.g., [8, 27]).

Sequencing technology can broadly be divided into three distinct generations:

i. First-generation technology, such as Sanger sequencing [15], was characterized by the use of fluorescently labeled chain-terminating dideoxy nucleotides, which were detected in sequencing machines.
ii. Second-generation sequencers (or high-throughput DNA next-generation sequencing; NGS) rely on shorter reads (between 75 and 400 bp) with higher error rates relative to the Sanger sequencing, but high sequence coverage (“massively parallel sequencing”). These methods generally use a solid support containing micro channels or wells in which sequencing by synthesis occurs. During synthesis, labeled nucleotides are incorporated, which produces signals allowing for imaging [20].

iii. Third-generation sequencers capable of analysing long DNA fragments rely on single-molecule real-time sequencing. They read longer stretches of DNA or RNA in a single pass, providing advantages such as longer read lengths and real-time data. A feature of third-generation sequencing technology is that the native DNA is sequenced directly without amplification. This has the advantage of removing nucleotide biases and alterations in relative abundance of DNA templates that are observed in some short-read sequence data [20].

In addition to the characterization of common variants typically included on SNP chips, use of whole-genome sequencing offers several supplemental benefits, including the characterization of rare variants and identification of other sources of variation, specifically structural and copy number variants. As mentioned by Mathieson and McVean [28], whole-genome sequencing supports studies of low frequency and rare variants, which are on average younger than common variants with high frequencies. Hence, minor polymorphisms are more powerful for distinguishing closely related populations and more informative with respect to recent demographic history [29, 30]. Copy number variants are a type of intermediate-scale structural variants with copy number changes involving a DNA fragment that is typically greater than 1 kb and less than 5 Mb. The importance of these repeats has been realized by their association to resistance/susceptibility to some diseases [31]. Structural variations are chromosomal rearrangements which include insertions, deletions, duplications, inversions, and translocations.

Next-generation sequencing coverage describes the average number of bases (i.e., read depth) that align to, or "cover," known reference bases.

Major factors that determine the required read depth in a genome sequencing study are the error rate of the sequencing method, the assembly algorithms used, the repeat complexity of the particular genome under study and the read length. Although increasing the sequencing coverage might add statistical power [32] or enable investigating very rare variants [33], it also increases the cost [34]. Therefore, low-pass sequencing has been proposed as a compromise between the cost of sequencing and the demand for millions of variants genotyped across the whole genome. It provides an affordable cost per sample, which makes it more accessible for large-scale studies [35]. Additionally, low-pass sequencing allows for high sample throughput, enabling the analysis of a large number of samples within a reasonable timeframe. Furthermore, it returns genotypes for both common and low-frequency variants, providing comprehensive genetic information [36]. However, low-pass sequencing has certain disadvantages. First, genotype calls obtained through low-pass sequencing may have low confidence due to the imputation process (i.e., predicting missing genetic data using known reference data), as direct genotyping is not possible, and genotypes are inferred [37]. Consequently, obtaining high-accuracy genotypes for specific loci is not possible, unlike in the case of array genotyping [36]. Furthermore, genotype calls are limited to variants present in the imputation reference panel [38]. These factors should be considered when utilizing low-pass sequencing for genotyping purposes.

Throughout this thesis, we explore several sequencing approaches in the Swiss Large White population that offer the optimal balance between cost and meeting the broad range of needs.

1.3 Imputation

Although the cost of genotyping has decreased by over 10,000-fold since 1990 [19], the number of markers and animals in-demand for genotyping is quickly growing. While the use of chips with higher SNP densities was shown to result in higher selection accuracies in genomic selection [39–41], the increased cost relative to low-density chips presented a barrier to widespread use. Much of this was resolved by the imputation methodology, where low-density genotypes are scaled up to high-density ones. Imputation to either higher array density [42] or to sequence level [43–45] has been applied in different livestock populations. As a result, genotyping strategies in large breeding schemes often involve the use of three or more densities of SNPs [46], which makes imputation for non-genotyped animals from genotyped relatives also possible [47].

Imputation of genotypes to the sequence level typically involves two steps. First, a phasing step that involves resolving haplotypes of high-density genotyped animals. Second, an imputation step where low-density genotypes are used together with linkage disequilibrium information to determine the combinations of haplotypes that are carried by animals not genotyped or genotyped at low-density. However, due to the lack of higher density genotypes available for a representative number of animals, direct imputation from medium-density genotypes to whole-genome sequence level had also been explored [48–50]. Imputation can be performed using several different programs (e.g., Beagle [51], Fimpute [52], AlphaImpute [47], STITCH [53], GLIMPSE1 [38] and GLIMPSE2 [54] etc.).

The accuracy of imputation is influenced by several factors, including the number of markers on the low-density genotyping panel, the

number of individuals that are genotyped at high-density, the local linkage disequilibrium (LD) between each low-density genotype and its surrounding high-density genotypes, and the number of high-density genotyped relatives of the individuals to be imputed [44, 55–57]. In pedigreed populations, the key factors that impact the accuracy of imputation are the density of genotypes of immediate ancestors and the density of genotypes in the reference panel used for reconstructing the haplotypes [47, 58]. This stresses the need of a densely genotyped reference population. Several population-based initiatives help to achieve a high-quality reference populations. The best-known large-scale resequencing initiative in livestock genomics is the 1000 Bull Genomes Project [59], though other large sequencing projects are being conducted by academic groups and breeding companies (PHARP [42], GENCOVE [35]). These efforts are largely inspired by the human 1,000 Genomes Project and hope to achieve deep characterization of the genetic variation between and within populations [60]. The availability of whole-genome sequence information for tens-to-hundreds of animals from specific breeds has proved useful to pinpoint the causative mutations underlying monogenic defects and facilitates fine-mapping and identification of causative variants for QTL detected by genome wide association studies (GWAS) [61, 62].

1.4 Genome annotation

Genome annotation is a crucial step in deciphering the genetic code of an organism. With a well-annotated genome, scientists can identify key regulatory regions and gene networks. Finding features of DNA is known as structural annotation [4] and includes annotating coding regions (open reading frames and genes) and non-coding regions (repeat sequence regions, pseudogenes, regulatory regions, etc.). The functional annotation

is a combination of the structural annotation to other forms of data that can be used to describe the products and roles of particular genomic elements [4, 63–65].

Gene refers to a genomic region that produces a polyadenylated mRNA, which encodes a protein. Eukaryotic genes often contain introns – ‘intervening sequences’ of bases that must be removed from the RNA transcript to make functional mRNA. Due to the presence of introns, gene prediction is a complex process. The annotation program reads DNA segments in all six possible reading frames (three reading frames on each of the two strands of DNA), searching for sequences of amino acids uninterrupted by 'stop' codons. A random sequence of triplets will include a stop codon approximately once in every 21 codons. If it finds a sequence of 240 bases without a stop codon, it is likely to contain genetic message. This sequence is called an open reading frame (ORF) and is considered a gene or a part of a gene [66]. Identifying genes is challenging, as an ORF may only represent a portion of a gene. The start and stop codons, promoter, and other regulatory elements can be separated from an ORF by multiple introns and exons [63, 66–68]. Additionally, the varying sizes of introns and alternative splicing further complicate gene structure prediction [69, 70].

Splicing is a crucial process where introns are removed from a pre-mRNA primary transcript and flanking exons are ligated [70]. Each intron contains three obligate spliceosome recognition signals: a 5′splice site, a branch site, and a 3′splice site [67, 71, 72]. Splicing occurs through a two-step mechanism. In the first step, the 2′-hydroxyl group of the branchpoint adenosine attacks the phosphodiester bond at the 5′splice site and displaces the 5′ exon while creating an RNA lariat; in the second step, the 3′-hydroxyl

group of the 5′ exon attacks the phosphodiester bond at the 3′ splice site and displaces the RNA lariat intron [68, 70, 72, 73]. This is catalyzed by a spliceosome, a complex of five small ribonucleic protein particles (U1, U2, U4, U5 and U6 snRNP) and other proteins [74, 75]. The motif encompassing the branch point has a consensus sequence 'nnyTrAy', where 'A' is the branchpoint residue, 'y' denotes pyrimidine (C or T), 'r' denotes purine (A or G), and 'n' denotes any nucleotide [73, 76–78]. The motif is recognized by U2 snRNP and undergoes imperfect base-pairing with its GUAGUA sequence [74, 79]. This imperfect interaction is characterized by a bulged nucleotide in the pre-mRNA that is activated as a branch point site [77]. In humans, nearly all protein-coding genes undergo splicing, and any disruption in this process can lead to the development of rare genetic diseases [80–84]; this emphasizes the significance of proper splicing for the accurate formation of functional proteins[82].

Exons can be spliced together differently, known as alternative splicing. This enables regulated generation of multiple mRNA and protein products from a single gene, and thus increases the diversity of the proteome [69]. For example, although the mouse and human genomes contain a similar number of genes, alternative pre-mRNA splicing occurs in >95% of human genes, compared to only ∼63% of mouse genes [69, 80, 85]. However, these numbers might be overestimated due to splicing errors, meaning the actual numbers of biologically relevant alternatively spliced genes are much lower [80].

Functional products of genes are RNAs and proteins. Genes that lead to the production of proteins are called protein-coding genes. Other genes that do not code proteins, but instead functional RNA molecules, are called noncoding genes. Noncoding RNA genes include genes for ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), small nuclear RNA

and nucleolar RNA (snRNA and snoRNA, respectively), and long noncoding RNA (lncRNA). The rest (~ 95%) of the genome is non-coding. That consists of non-functional "junk DNA" and potentially biologically meaningful regions, such as regulatory elements, non-coding genes, introns, pseudogenes, or transposable elements [86].

1.5 Genetic diversity

Genetic diversity is a commonly used term referring to the set of differences between individuals, breeds, and species in their DNA. Variable alleles and haplotypes are important parts of this genetic diversity and can be detected more precisely with the new genomic tools. Reliable sets of genotypes obtained by whole-genome sequencing or through imputation provide useful insights into variation within or between populations. This information can be used to reveal population history of bottlenecks, admixture, migration, estimation of current genetic diversity, and links to phenotypic variation. For instance, diversity attributed to variable patterns of domestication and geographic distributions of livestock revealed differences between populations and signatures of selection, which assisted with the identification candidate genes (e.g.,[87]).

Genetic diversity is essential for populations to respond to environmental changes, with implications in terms of, for example, human health, breeding strategies in crops and farm animals, management of infectious diseases and conservation of endangered species. Natural populations of sufficiently interbred animals can gradually adapt to the specific conditions in which the population lives, e.g., high altitude, rough climate, or tropical conditions. In 1930, Fisher [88] formulated the following theorem: "The rate of increase in fitness of any organism at any

time is equal to its genetic variance in fitness at that time". This aspect of natural selection for adaptive traits has a new importance from the perspective of climate change because agricultural species will need to adapt to increasingly rapid changes in environmental conditionals [89]. The genetic diversity differs considerably between species as well as between chromosomes [90].

Prior to the investigation of DNA, it was only known that there is a 50% chance that two offspring would inherit the same allele from the parent, and on average two offspring would share 50% of their DNA. The outcome of these chance events during meiosis is that two full siblings may share substantially more or substantially less than 50% of inherited DNA [91, 92]. With the possibility of genotyping, genetic differences and similarities between siblings can be established soon after conception.

The genetic diversity within domesticated species is primarily found among the various breeds that have been developed. 'Breed' can be defined as a subgroup of a species with a common history whose members are treated in a common manner with respect to genetic management [93]. In genomics, the breed can be defined based on sufficiently dense SNP chips with a range of techniques (e.g., principal components or multi-dimensional scaling), which separate breeds into discrete clusters displayed in two or three dimensions. These analyses can go beyond this and separate breeds into sub-groups with limited gene flow between each other. In such approaches the use of diverse samples is desirable in order to gain the perspective in clustering the animals. It is questionable whether all breeds significantly contribute to the genetic diversity of their species and consequently whether all these breeds must be included in utilization schemes (breeding programs) to maintain the within-breed diversity, or if

all breeds should be conserved in national conservation plans or gene banks.

The "key ancestor approach" refers to a method in genetic studies where a subset of individuals from a population, known as "key ancestors," is selected for genomic sequencing. These key ancestors are chosen strategically to represent the genetic diversity and variability present in the entire population [94].

1.5.1 Measures of diversity

Different molecular measures of genetic diversity are used in literature: percentage of polymorphic sites, distribution of allelic frequencies, expected heterozygosity, observed heterozygosity, and allelic diversity [93]. The genetic diversity can be found within breeds as well as between breeds, i.e., deviations from breed mean or between different breed means.

In population genetics, the most used measure of differentiation between populations is the Fixation Index of Wright (F_{ST}) [95]. It can be defined as the proportion of total diversity that appears between breeds. It provides a simple way of calculating the contribution of each breed to the total genetic diversity in the species.

A measure of genetic diversity within populations is nucleotide diversity (π) [96], or the average proportion of nucleotides that differ between any randomly sampled pair of sequences. This measure incorporates information regarding the extent of differentiation between sequences as well the relative frequencies of the sequences in the sample. Nucleotide diversity is similar to the classic measure of heterozygosity and is not greatly influenced by rare alleles. Another within-population

parameter is runs of homozygosity (ROH), long haplotypes within the same individual that are identical (i.e., homozygous for all the loci within) [97, 98]. Detection of these segments is performed in 'sliding windows', or predefined blocks moved along the genome. The proportion of the genome included in ROH is a measure of inbreeding (F_{ROH}) [99, 100].

ROH mainly reflect identity by descent (IBD) [98], as it is highly unlikely to carry two identical long haplotypes if they are not copies of an ancestral one. Therefore, the expectation is that long ROH comes from a recent ancestor and reflects recent inbreeding while shorter ones are from more distant ancestors [98, 100, 101]. However, the precise distribution of the length of ROH is still poorly understood and its interpretation varies considerably between studies [102–104]. This is due to the different parameters being considered for ROH, specifically SNPs, the minimum density SNPs within the segment, the number of missing genotypes allowed, and the number of heterozygous positions allowed (to allow for genotyping errors). Ferenčaković et al. [100] performed a sensitivity analysis using genomic data on three cattle breeds and showed how the appropriate length for ROH is also a function of the coverage of the SNP chip used, with less dense panels detecting false ROH when pursuing for short segments.

1.5.2 Population demographic

Demographic events — such as bottlenecks, migrations, admixture, and selected crossbreeding — have largely shaped the genome of domesticated species and contributed to considerable diversity among populations [105–109]. During these events the number of breeding individuals in a population is changing, e.g., is dramatically reduced in bottlenecks, or genetically diverged from its original source population during admixture. Reducing the species to a few hundred greatly and

rapidly reduces the number of rare alleles, and thus, the genetic diversity of population. This has been observed during the domestication of cattle [22], and in milder form in sheep [110] and horse [111].

Selection acts in three modes (positive, negative, and balancing) during which one or more alleles are favored or disfavored [19, 112–116]. Each mode leaves a specific signature on genomic variation and architecture, while the level of evolutionary constraint is a function of effective population size. For example, deleterious allele, which was not meant to be tolerated, drifts during bottleneck to a high frequency [117].

Signatures of selection can be investigated at the population level or across breeds to highlight potentially relevant functional polymorphisms [118–122]. Positive selection on the DNA reduces local variability, and through hitchhiking, the effect increases homozygosity in regions [115]. This has been considered in several measures: ROH (viz. above); 'extended haplotype homozygosity', which detects selection signatures by comparing a high frequency and extended homozygosity based haplotype with other haplotypes at the selected locus [115]or in 'integrated haplotype score', which is a measure of the amount of extended haplotype homozygosity at a given SNP along the ancestral allele relative to the derived allele, typically standardized to the distribution of observed iHS scores over a range SNPs with similar derived allele frequencies [123].

Several statistical tests, which are based on predicted effects relative to the standard neutral model, have been proposed for inferring "selective sweeps". These include (1) an excess of rare alleles compared to the standard neutral model with Tajima's D [124] and (2) method to test the significance of a local reduction of variation and a skew of the frequency

spectrum caused by a hitchhiking event using ‘composite likelihood ratio’ test [125, 126].

Possible validation of selection scans is combining the selection analysis and GWAS. Several studies demonstrated that genomic regions that exhibit selection signatures are also enriched for genes associated with biologically important traits [118, 122, 127]. Both estimators rely on the underlying LD between the causal variant and the genotyped SNP. If regions showing evidence of positive selection are present in the QTL regions, they would be excellent candidates for containing the causative alleles influencing the trait of interest. For example, Qanbari et al. [128] showed a perfect overlap between selection candidates for appearance traits with major coat color QTLs in cattle. However, the co-localization of a selective sweep and a QTL for a relevant phenotype does not necessarily mean that the two signals are correlated and might require very special biological conditions.

LD is the non-random association of alleles at two separate loci within a population [129]. LD patterns across the genome can be influenced by various evolutionary forces, including migration, mutation, genetic drift, natural selection, population structure, and recombination rates [130]. Consequently, LD maps serve as valuable tools for studying genetic diversity, identifying selective sweeps in livestock populations [131], and estimating effective population sizes [104]. The most commonly used measure is the squared correlation coefficient (r^2), which quantifies the strength of LD.

1.6 Genomic selection

Genomic selection is a quantitative genetic method that utilizes molecular markers to predict an individual's breeding value for a trait of interest [132]. It has been extensively applied in various agricultural species, including livestock [133] and crops [134].

Traits used in pig breeding are typically chosen based on a number of factors, including the requirements of the pork industry and the needs of individual producers [135]. Quantitative traits, such as growth rate and feed efficiency, are often prioritized for their economic impact, while qualitative traits, such as meat quality and disease resistance, may also be important for meeting consumer demand and ensuring animal welfare [19].

Selection of animals involves ranking according to estimated breeding value. It is calculated by analyzing the animal's own performance data, as well as performance data from its relatives and represents the portion of an animal's genetic makeup that can be passed on to its offspring and used to improve the performance of the next generation [19, 47].

Since the development of best linear unbiased prediction (BLUP) methodology [136] up to recent advanced methods for genomic evaluation [137], the estimation of breeding values relies on an additive genetic relationship matrix. It exploits information on relatives to account for the genetic (co)variance structure among individuals in the population. The relationship matrix contains coefficients of relationship that reflect kinship between the individuals by calculating the probability of sharing identical alleles by descent. This measure, as defined by Falconer and Mackay [138], determines the additive genetic relationship between two individuals, which is double their kinship.

Marker-assisted selection was first applied in commercial pig breeding in the early 1990s when the Hal-1843 marker test became available for selection against a mutation in the *RYR1* gene that lead to poor meat quality in stressful conditions [8]. By the early 2000s, quantitative geneticists had proposed new methods for developing the marker-assisted selection, which required genotyping multiple markers across the genome to identify any unknown loci with a causative effect on the phenotype [132]. This development led to the demand for SNPs to be incorporated into a chip for efficient genotyping of multiple sets of markers and estimation of genomic breeding values based on cumulative SNP effects [139]. The advancements in livestock genomics have resulted in a greater understanding of the genetic architecture of livestock species. These technologies are now used in almost all major livestock species in developed countries [140].

1.7 Swiss pig breeding

Commercial Swiss pig breeding follows a pyramidal structure with three units (nucleus, multiplication, and production). Approximately 10,000 sows and 500 boars are bred in the nucleus unit to enable the genetic improvement of paternal and maternal breeds. The purebred animals from two maternal breeds are crossed in the multiplication unit to produce fertile and resilient gilts. These crossbred sows are eventually inseminated with terminal boars from paternal breeds in the production unit to produce roughly 2.5 million fattening pigs every year.

One of the main breeds used in the core breeding program is Swiss Large White (SLW). In the early 2000s, animals with desirable traits were selected from this initially purely maternal breed to form the base of a sire

line. Further enhancements continued for several generations until a closed breeding population was established roughly ten generations ago.

1.8 References

1. A reference standard for genome biology. Nat Biotechnol. 2018;36:1121–1121.
2. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009;4.
3. Hu ZL, Park CA, Reecy JM. Developmental progress and current status of the Animal QTLdb. Nucleic Acids Res. 2016;44 Database issue:D827.
4. García-Sancho M, Lowe J. Making Reference Genomes Useful: Annotation. 2023;:205–54.
5. Schook LB, Beever JE, Rogers J, Humphray S, Archibald A, Chardon P, et al. Swine Genome Sequencing Consortium (SGSC): A Strategic Roadmap for Sequencing The Pig Genome. Comp Funct Genomics. 2005;6:251.
6. Robert C, Fuentes-Utrilla P, Troup K, Loecherbach J, Turner F, Talbot R, et al. Design and development of exome capture sequencing for the domestic pig (Sus scrofa). BMC Genomics. 2014;15:1–9.
7. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience. 2020;9:1–14.
8. Fujii J, Otsu K, Zorzato F, De Leon S, Khanna VK, Weiler JE, et al. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science (1979). 1991;253:448–51.
9. Hayes BJ, Lewin HA, Goddard ME. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends Genet. 2013;29:206–14.
10. Van Eenennaam AL, Young AE. Prevalence and impacts of genetically engineered feedstuffs on livestock populations. J Anim Sci. 2014;92:4255–78.

11. Berry DP, Conroy S, Pabiou T, Cromie AR. Animal breeding strategies can improve meat quality attributes within entire populations. Meat Sci. 2017;132:6–18.

12. Zak LJ, Gaustad AH, Bolarin A, Broekhuijse MLWJ, Walling GA, Knol EF. Genetic control of complex traits, with a focus on reproduction in pigs. Mol Reprod Dev. 2017;84:1004–11.

13. Dekkers J, Rowland RRR, Lunney JK, Plastow G. Host genetics of response to porcine reproductive and respiratory syndrome in nursery pigs. Vet Microbiol. 2017;209:107–13.

14. Tortereau F, Servin B, Frantz L, Megens HJ, Milan D, Rohrer G, et al. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics. 2012;13:586.

15. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94.

16. Wiedmann RT, Smith TPL, Nonneman DJ. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. 2008;9.

17. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods. 2008;5:247–52.

18. Nonneman DJ, Lents CA. Functional genomics of reproduction in pigs: Are we there yet? Mol Reprod Dev. 2022. https://doi.org/10.1002/MRD.23625.

19. Knol EF, Nielsen B, Knap PW. Genomic selection in commercial pig breeding. Animal Frontiers. 2016;6:15–22.

20. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. Journal of Applied Genetics 2011 52:4. 2011;52:413–35.

21. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005;15:1496–502.

22. MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, Goddard ME. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. Mol Biol Evol. 2013;30:2209.

23. Groenen MAM. A decade of pig genome sequencing: A window on pig domestication and evolution. Genetics Selection Evolution. 2016;48:1–9.

24. Grady DL, Ratliff RL, Robinson DL, Mccanlies EC, Meyne J, Moyzis RK. Highly conserved repetitive DNA sequences are present at human centromeres. Proc Natl Acad Sci U S A. 1992;89:1695–9.

25. Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 2012;491:393–8.

26. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443–8.

27. Winter A, Krämer W, Werner FAO, Kollers S, Kata S, Durstewitz G, et al. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:Diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proc Natl Acad Sci U S A. 2002;99:9300–5.

28. Mathieson I, McVean G. Demography and the Age of Rare Variants. PLoS Genet. 2014;10.

29. Cai Z, Sarup P, Ostersen T, Nielsen B, Fredholm M, Karlskov-mortensen P, et al. Animal Genetics and Genomics Genomic diversity revealed by whole-genome sequencing in three Danish commercial pig breeds. 2020;98:1–12.

30. Barendse W, Harrison BE, Bunch RJ, Thomas MB, Turner LB, W. B, et al. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. BMC Genomics. 2009;10:178.

31. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, et al. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science (1979). 2009;324:522–8.

32. Huang Y, Hickey JM, Cleveland MA, Maltecca C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet Sel Evol. 2012;44:25.

33. Ros-Freixedes R, Valente BD, Chen C-Y, Herring WO, Gorjanc G, Hickey JM, et al. Rare and population-specific functional variation across pig lines. Genetics Selection Evolution. 2022;54:1–16.

34. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 2014 15:2. 2014;15:121–32.

35. Data analysis configurations - Gencove Docs. https://docs.gencove.com/main/data-analysis-configurations/. Accessed 1 Jul 2023.

36. Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, et al. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. 2021;22:197.

37. Chat V, Ferguson R, Morales L, Kirchhoff T. Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies. Front Genet. 2022;12:2712.

38. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53:120–6.

39. Zhu B, Zhang J-J, Niu H, Long G, Peng G, Xu L-Y, et al. Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. J Integr Agric. 2017;2017:911–20.

40. Erbe M, Gredler B, Seefried FR, Simianer H. A Function Accounting for Training Set Size and Marker Density to Model the Average Accuracy of Genomic Prediction. 2013. https://doi.org/10.1371/journal.pone.0081046.

41. Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. Comparison of genomic predictions using medium-density (∼54,000) and high-density (∼777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci. 2012;95:4657–65.

42. Wang Z, Zhang Z, Chen Z, Sun J, Cao C, Wu F, et al. PHARP: A pig haplotype reference panel for genotype imputation. bioRxiv. 2021;:2021.06.03.446888.

43. Cai Z, Christensen OF, Lund MS, Ostersen T, Sahana G. Large-scale association study on daily weight gain in pigs reveals overlap of genetic factors for growth in humans. BMC Genomics. 2022;23:1–13.

44. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genetics Selection Evolution. 2017;49.

45. Yan G, Qiao R, Zhang F, Xin W, Xiao S, Huang T, et al. Imputation-Based Whole-Genome Sequence Association Study Rediscovered the Missing QTL for Lumbar Number in Sutai Pigs. Sci Rep. 2017;7:1–10.

46. Ventura R V., Miller SP, Dodds KG, Auvray B, Lee M, Bixley M, et al. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genetics Selection Evolution. 2016;48:1–20.

47. Hickey JM, Kinghorn BP, Tier B, Van Der Werf JH, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics Selection Evolution. 2012;44:1–11.

48. Casu S, Usai MG, Sechi T, Salaris SL, Miari S, Mulas G, et al. Association analysis and functional annotation of imputed sequence data within genomic regions influencing resistance to gastro-intestinal parasites detected by an LDLA approach in a nucleus flock of Sarda dairy sheep. Genetics Selection Evolution. 2022;54:1–16.

49. Van Den Berg S, Vandenplas J, Van Eeuwijk FA, Bouwman AC, Lopes MS, Veerkamp RF. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. Genetics Selection Evolution. 2019;51.

50. Yan G, Liu X, Xiao S, Xin W, Xu W, Li Y, et al. An imputed whole-genome sequence-based GWAS approach pinpoints causal mutations for complex traits in a specific swine population. Sci China Life Sci. 2021;:1–14.

51. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103:338–48.

52. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15.

53. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nat Genet. 2016;48:965.

54. Biobank UK, Rubinacci S, Hofmeister R, Sousa Da Mota B, Delaneau O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. bioRxiv. 2022;:2022.11.28.518213.

55. Da Costa Hermisdorff I, Bermal Costa R, Galvão De Albuquerque L, Pausch H, Kadri NK. Investigating the accuracy of imputing autosomal

variants in Nellore cattle using the ARS-UCD1.2 assembly of the bovine genome. BMC Genomics. 2020;:21–772.

56. Lloret-Villas A, Pausch H, Leonard AS. Size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle. bioRxiv. 2023. https://doi.org/10.1101/2023.01.13.523894.

57. Bolormaa S, Chamberlain AJ, Khansefid M, Stothard P, Swan AA, Mason B, et al. Accuracy of imputation to whole-genome sequence in sheep 06 Biological Sciences 0604 Genetics. Genetics Selection Evolution. 2019;51:1–17.

58. Zhang Z, Ma P, Zhang Z, Wang Z, Wang Q, Pan Y. The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in Duroc pigs. Genomics. 2022;114:340–50.

59. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. 2018. https://doi.org/10.1146/annurev-animal-020518.

60. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nature Reviews Genetics 2018 20:3. 2018;20:135–56.

61. Wang Y, Xiao L, Guo S, An F, Du D. Fine Mapping and Whole-Genome Resequencing Identify the Seed Coat Color Gene in Brassica rapa. PLoS One. 2016;11:e0166464.

62. Cáceres G, López ME, Cádiz MI, Yoshida GM, Jedlicki A, Palma-Véjares R, et al. Fine Mapping Using Whole-Genome Sequencing Confirms Anti-Müllerian Hormone as a Major Gene for Sex Determination in Farmed Nile Tilapia (Oreochromis niloticus L.). G3 Genes|Genomes|Genetics. 2019;9:3213–23.

63. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet. 2010;11:559–71.

64. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 2019;20:1–3.

65. Karlsson M, Sjöstedt E, Oksvold P, Sivertsson Å, Huang J, Álvez MB, et al. Genome-wide annotation of protein-coding genes in pig. BMC Biol. 2022;20:1–18.

66. Susman M. Genes: Definition and Structure. eLS. 2014. https://doi.org/10.1002/9780470015902.A0001494.PUB3.

67. Jo B-S, Choi SS. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. 2015;13:112.

68. Keller EB, Noon WA. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc Natl Acad Sci U S A. 1984;81:7417.

69. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science (1979). 2012;338:1587–93.

70. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem. 2015;84:291.

71. Zhang X, Zhang Y, Wang T, Li Z, Cheng J, Ge H, et al. A Comprehensive Map of Intron Branchpoints and Lariat RNAs in Plants. Plant Cell. 2019;31:956.

72. Berget SM. Exon recognition in vertebrate splicing. J Biol Chem. 1995;270:2411–4.

73. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. Genome Res. 2015;25:290.

74. Lamond AI. The Spliceosome. BioEssays. 1993;15:595–603.

75. Buratti E, Baralle FE. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Mol Cell Biol. 2004;24:10505.

76. Signal B, Gloss BS, Dinger ME, Mercer TR. Machine learning annotation of human branchpoints. Bioinformatics. 2018;34:920–7.

77. Taggart AJ, Lin CL, Shrestha B, Heintzelman C, Kim S, Fairbrother WG. Large-scale analysis of branchpoint usage across species and cell lines. Genome Res. 2017;27:639–49.

78. Gao K, Masuda A, Matsuura T, Ohno K. Human branch point consensus sequence is yUnAy. Nucleic Acids Res. 2008;36:2257.

79. Zhuang Y, Goldstein AM, Weiner AM. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proc Natl Acad Sci U S A. 1989;86:2752–6.

80. Bhuiyan SA, Ly S, Phan M, Huntington B, Hogan E, Liu CC, et al. Systematic evaluation of isoform function in literature reports of alternative splicing. BMC Genomics. 2018;19.

81. Shiau C-K, Huang J-H, Liu Y-T, Tsai H-K. Genome-wide identification of associations between enhancer and alternative splicing in human and mouse. BMC Genomics. 2022;22:1–12.

82. Sanders SJ, Schwartz GB, Farh KKH. Clinical impact of splicing in neurodevelopmental disorders. Genome Med. 2020;12:1–5.

83. Blakes AJM, Wai HA, Davies I, Moledina HE, Ruiz A, Thomas T, et al. A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. Genome Med. 2022;14:1–11.

84. Nosková A, Hiltpold M, Janett F, Echtermann T, Fang ZH, Sidler X, et al. Infertility due to defective sperm flagella caused by an intronic deletion in DNAH17 that perturbs splicing. Genetics. 2021;217.

85. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338:1593–9.

86. Fan K, Pfister E, Weng Z. Toward a comprehensive catalog of regulatory elements. Hum Genet. 2023;:1–21.

87. Yang J, Li W-R, Lv F-H, He S-G, Tian S-L, Peng W-F, et al. Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. Mol Biol Evol. 2016;33:2576–92.

88. Fisher R. The distribution of gene ratios for rare mutations. Proc Roy Soc Edinb. 1930;50:205–20.

89. Hoffmann I. Adaptation to climate change-exploring the potential of locally adapted breeds. Animal. 2013;7:346–62.

90. Ellegren H, Galtier N. Determinants of genetic diversity. Nature Reviews Genetics 2016 17:7. 2016;17:422–33.

91. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2006;2:0316–25.

92. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009;91:47–60.

93. Oldenbroek K, Waaij L van der. Textbook animal breeding : animal breeding and genetics for BSc students. 2014.

94. Goddard ME, Hayes BJ. Genomic Selection Based on Dense Genotypes Inferred From Sparse Genotypes. Proc Assoc Advmt Anim Breed Genet. 2009;18.

95. Wright S. The genetical structure of populations. Ann Eugen. 1951;15:323–54.

96. Wright S. Testing for effects of recombination rate on nucleotide diversity in natural populations of Arabidopsis lyrata. Genetics. 2006;174:1421–30.

97. Broman KW, Weber JL. Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. Am J Hum Genet. 1999;65:1493.

98. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: Windows into population history and trait architecture. Nature Reviews Genetics. 2018;19:220–34.

99. Purfield DC, Berry DP, McParland S, Bradley DG. Runs of homozygosity and population history in cattle. BMC Genet. 2012;13:1–11.

100. Ferenčaković M, Hamzić E, Gredler B, Solberg TR, Klemetsdal G, Curik I, et al. Estimates of autozygosity derived from runs of homozygosity: Empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics. 2013;130:286–93.

101. Kirin M, McQuillan R, Franklin CS, Campbell H, Mckeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS One. 2010;5:e13996.

102. Bhati M, Kadri NK, Crysnanto D, Pausch H. Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. BMC Genomics. 2020;21:1–14.

103. Xie R, Shi L, Liu J, Deng T, Wang L, Liu Y, et al. Genome-wide scan for runs of homozygosity identifies candidate genes in three pig breeds. Animals. 2019;9:1–14.

104. Shi L, Wang L, Liu J, Deng T, Yan H, Zhang L, et al. Estimation of inbreeding and identification of regions under heavy selection based on runs of homozygosity in a Large White pig population. J Anim Sci Biotechnol. 2020;11:1–10.

105. Andersson L, Georges M. Domestic-animal genomics: Deciphering the genetics of complex traits. Nature Reviews Genetics. 2004;5:202–12.

106. Giuffra E, Kijas JMH, Amarger V, Carlborg Ö, Jeon JT, Andersson L. The origin of the domestic pig: Independent domestication and subsequent introgression. Genetics. 2000;154:1785–91.

107. Frantz LAF, Schraiber JG, Madsen O, Megens HJ, Cagan A, Bosse M, et al. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nat Genet. 2015;47:1141–8.

108. Wright D. The genetic architecture of domestication in animals. Bioinform Biol Insights. 2015;9:11–20.

109. Ramos-Onsins SE, Burgos-Paz W, Manunza A, Amills M. Mining the pig genome to investigate the domestication process. Heredity (Edinb). 2014;113:471–84.

110. Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LR, Cristobal MS, et al. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol. 2012;10:e1001258.

111. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009;326:865–7.

112. Nielsen R. Molecular signatures of natural selection. Annual Review of Genetics. 2005;39:197–218.

113. Biswas S, Akey JM. Genomic insights into positive selection. Trends in Genetics. 2006;22:437–46.

114. Fisher R. The Genetical Theory of Natural Selection. 1999.

115. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312:1614–20.

116. Dukler N, Mughal MR, Ramani R, Huang YF, Siepel A. Extreme purifying selection against point mutations in the human genome. Nat Commun. 2022;13:1–12.

117. Marsden CD, Vecchyo DO Del, O'Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci U S A. 2016;113:152–7.

118. Rubin C-JJ, Megens H-JJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109:19529–36.

119. Qanbari S, Simianer H. Mapping signatures of positive selection in the genome of livestock. Livest Sci. 2014;166:133–43.

120. Xu L, Bickhart DM, Cole JB, Schroeder SG, Song J, Van Tassell CP, et al. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. Mol Biol Evol. 2015;32:711–25.

121. Sahm A, Bens M, Szafranski K, Holtze S, Groth M, Görlach M, et al. Long-lived rodents reveal signatures of positive selection in genes associated with lifespan. PLoS Genet. 2018;14.

122. Zhang W, Yang M, Zhou M, Wang Y, Wu X, Zhang X, et al. Identification of Signatures of Selection by Whole-Genome Resequencing of a Chinese Native Pig. Front Genet. 2020;11:566255.

123. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:0446–58.

124. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.

125. Besag J. Statistical Analysis of Non-Lattice Data. The Statistician. 1975;24:179.

126. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002;160:765.

127. Howard DM, Pong-Wong R, Knap PW, Woolliams JA. Use of haplotypes to identify regions harbouring lethal recessive variants in pigs. Genetics Selection Evolution. 2017;49:1–10.

128. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genet. 2014;10:e1004148.

129. Weir BS. Inferences about Linkage Disequilibrium. Biometrics. 1979;35:235.

130. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet. 2002;3:299–309.

131. Gutiérrez-Gil B, Arranz JJ, Wiener P. An interpretive review of selective sweep studies in Bos taurus cattle populations: identification

of unique and shared selection signals across breeds. Front Genet. 2015;6 MAY.

132. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819.

133. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.

134. Heffner EL, Sorrells ME, Jannink JL. Genomic Selection for Crop Improvement. Crop Sci. 2009;49:1–12.

135. Knap PW. Pig breeding goals in competitive markets. Proceedings, 10 th World Congress of Genetics Applied to Livestock Production. 2014;:17–22.

136. Henderson CR, Kempthorne O, Searle SR, von Krosigk CM. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. Biometrics. 1959;15:192.

137. Fernando RL, Cheng H, Garrick DJ. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. Genetics Selection Evolution. 2016;48:1–12.

138. Falconer DS, Mackay TFC. Introduction to quantitative genetics. Essex. UK: Longman Group. 1996;:448.

139. Knol EF, Nielsen B, Knap PW. Genomic selection in commercial pig breeding. Animal Frontiers. 2016;6:15–22.

140. Saravanan KA, Panigrahi M, Kumar H, Nayak SS, Rajawat D, Bhushan B, et al. Progress and future perspectives of livestock genomics in India: a mini review. Anim Biotechnol. 2022. https://doi.org/10.1080/10495398.2022.2056046.

2 Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs

Adéla Nosková[1], Meenu Bhati[1], Naveen Kumar Kadri[1], Danang Crysnanto[1], Stefan Neuenschwander[2], Andreas Hofer[3], Hubert Pausch[1]

[1]Animal Genomics, ETH Zürich, Eschikon 27, 8315 Lindau, Switzerland

[2]Animal Genetics, ETH Zürich, Tannenstrasse 1, 8092 Zürich, Switzerland

[3]SUISAG, Allmend 8, 6204 Sempach, Switzerland

Published in *BMC Genomics, (2021), 290, 22(1).*

https://doi.org/10.1186/s12864-021-07610-5

Contribution: I participated in conceiving the study, analysing the results, and writing of the manuscript.

2.1 Abstract

2.1.1 Background

The key-ancestor approach has been frequently applied to prioritize individuals for whole-genome sequencing based on their marginal genetic contribution to current populations. Using this approach, we selected 70 key ancestors from two lines of the Swiss Large White breed that have been selected divergently for fertility and fattening traits and sequenced their genomes with short paired-end reads.

2.1.2 Results

Using pedigree records, we estimated the effective population size of the dam and sire line to 72 and 44, respectively. In order to assess sequence variation in both lines, we sequenced the genomes of 70 boars at an average coverage of 16.69-fold. The boars explained 87.95 and 95.35% of the genetic diversity of the breeding populations of the dam and sire line, respectively. Reference-guided variant discovery using the GATK revealed 26,862,369 polymorphic sites. Principal component, admixture and fixation index (F_{ST}) analyses indicated considerable genetic differentiation between the lines. Genomic inbreeding quantified using runs of homozygosity was higher in the sire than dam line (0.28 vs 0.26). Using two complementary approaches, we detected 51 signatures of selection. However, only six signatures of selection overlapped between both lines. We used the sequenced haplotypes of the 70 key ancestors as a reference panel to call 22,618,811 genotypes in 175 pigs that had been sequenced at very low coverage (1.11-fold) using the GLIMPSE software. The genotype concordance, non-reference sensitivity and non-reference discrepancy between thus inferred and Illumina PorcineSNP60 BeadChip-called genotypes was 97.60, 98.73 and 3.24%, respectively. The low-pass

sequencing-derived genomic relationship coefficients were highly correlated ($r > 0.99$) with those obtained from microarray genotyping.

2.1.3 Conclusions

We assessed genetic diversity within and between two lines of the Swiss Large White pig breed. Our analyses revealed considerable differentiation, even though the split into two populations occurred only few generations ago. The sequenced haplotypes of the key ancestor animals enabled us to implement genotyping by low-pass sequencing which offers an intriguing cost-effective approach to increase the variant density over current array-based genotyping by more than 350-fold.

2.2 Background

Swine production follows the classical breeding pyramid. Genetic gain is generated in nucleus herds and transmitted via the multiplier to the production unit. Swiss pig production relies on maternal and paternal Swiss Large White (SLW) lines at the top level of the breeding pyramid. For decades, the SLW breed has been maintained as a universal breed, selected for production and fertility traits. In 2002, the population was divided into sire and dam lines that have been divergently selected for fattening and reproduction since then. Approximately 32.5% and 30% of the genes of 2.5 million fattening pigs slaughtered in 2020 in Switzerland originate from the dam and sire line, respectively [1]. Both lines are maintained in purebred nucleus herds. However, little is known about the genetic diversity within the lines.

The SLW breeding boars are selected based on genome-based breeding values that are predicted using genotypes obtained with a

customized version of the Illumina PorcineSNP60 BeadChip. Apart from a small number of putatively causal variants that are included in the custom part, the content of the currently used microarray was designed in a way that it is useful for mainstream breeds [2]. However, the genetic constitution of the SLW breed beyond the microarray-derived SNP remains largely unknown. The sequencing of key ancestor animals has been proposed as a cost-efficient way to assess sequence variation within a population. The genomes of key ancestor individuals maximally represent the genetic diversity of the target population [3, 4]. Due to the use of individual boars in artificial insemination and intense selection in nucleus herds, the effective population size of most pig breeding populations is low. Thus, most common polymorphic sites segregating in the population can be traced back to the genomes of important contributors to the current population [5, 6]. The key ancestor approach was frequently applied to identify the most important contributors to current cattle breeding populations [6]. Recently it was also used to prioritize animals for sequencing in commercial pig breeding lines [7].

The availability of sequence variant genotypes from key ancestor animals enables imputing sequence-level genotypes for animals that had been genotyped at lower density [8–10]. In livestock populations that are routinely genotyped using 60K genotyping arrays, sequence variant genotypes are typically imputed using stepwise imputation [11]. In a first step, 60K genotypes are imputed to higher density (e.g., 700K) using animals that have been genotyped with high-density genotyping arrays. In a second step, the partially imputed high-density genotypes are imputed to the sequence level based on a sequenced reference panel. The accuracy of imputing 60K genotypes directly to the sequence level is low, particularly for rare variants, rendering most of them uninformative for downstream

analyses such as genomic prediction and association testing [12, 13]. Reference-guided variant phasing and imputation from low-pass sequencing data offers an intriguing alternative approach to the two-step imputation approach in pedigreed populations [14]. This approach utilises a sequenced haplotype reference panel that represents the diversity of the target population. Sequence variant genotypes of animals sequenced at very shallow coverage are then inferred conditional on the observed haplotypes of the reference panel. This method is particularly useful in species for which dense microarray-derived genotypes are not available. Recent investigations [8, 15, 16] suggest that a sequencing coverage less than 1-fold is sufficient to accurately infer genotypes at known loci - provided an informative haplotype reference panel is available.

Here we obtain whole-genome sequencing data from key ancestor animals to characterize genetic diversity, population structure, and signatures of selection in two divergently selected commercial pig breeds. Using the haplotypes of the key ancestor animals as a reference panel, we accurately genotype more than 22 million variants in animals that have been sequenced at low coverage.

2.3 Results

Using pedigree records, the average inbreeding coefficients of the active breeding animals of the sire and dam line were 0.06 ± 0.02 and 0.05 ± 0.01, respectively. Based on these values and the inbreeding coefficients of the parents, we estimated the effective population size of the sire and dam line of the Swiss Large White (SLW) breed to 44 and 72, respectively. In order to assess sequence variation within the two lines, we prioritized 70 boars for whole-genome sequencing based on their marginal genetic

contributions to the active breeding populations with a key ancestor approach. Of the 70 boars, 38 and 32 represent the sire and dam line, respectively, explaining 95.35 and 87.95% of the genetic diversity of the active breeding populations.

Following quality control (removal of adapter sequences, reads and bases of low sequencing quality), between 81.15 and 377.01 million read pairs (2 x 150 bp) per sample (mean: 165.55 ± 60.32 million read pairs) were aligned to the SSC11.1 assembly of the porcine genome. Using reads with high mapping quality (reads with mapping quality < 10 and SAM bitwise flag 1796 were not considered), the average sequencing coverage of the 70 boars was 16.69 ± 5.93-fold across all autosomes. Raw sequence read data of 70 pigs have been deposited at the European Nucleotide Archive (ENA) of the EMBL at BioProject PRJEB38156 and PRJEB39374.

A reference-guided multi-sample variant discovery and genotyping approach yielded genotypes at 28,407,060 sites (22,191,375 biallelic SNP, 4,379,470 biallelic INDEL, and 1,836,215 others, Table 2.1). We applied GATK's VariantFiltration module for site-level hard filtration using parameters recommended in the best practice guidelines [17]. Subsequently, we applied Beagle (version 4.1; [18]) phasing and imputation to improve the genotype calls from GATK and to impute sporadically missing genotypes. Following the imputation, we retained 26,862,369 variants including 21,592,583 SNP and 5,269,786 INDEL. The number of polymorphic sites that were seen in the heterozygous (singletons) and homozygous (doubletons) state only once was 2,026,088 (7.54%) and 72,100 (0.27%), respectively. To prevent bias resulting from flawed genotypes in repetitive regions, we excluded 1,710,337 variants for which an excess of sequencing coverage was evident for downstream

analyses. The transition/transversion (Ti/Tv)-ratio estimated from filtered and imputed variants was 2.28.

The resulting data were separated into two datasets containing 23,774,053 and 23,531,919 autosomal variants detected in 32 and 38 boars from the dam and sire line, respectively. Of the variants, 1,049,689 and 1,594,775 were fixed for the alternate allele in the dam and sire line, respectively. On average, we detected 11,119,760 ± 176,113 biallelic variants per animal (Figure 2.1A), of which 6,258,456 ± 280,127 and 4,861,304 ± 135,524 were heterozygous and homozygous for the reference allele, respectively. The average nucleotide diversity (π) across 452,444 overlapping windows (10 kb in size with 5 kb steps), spanning 22,840,217 and 22,529,446 biallelic variants, respectively, was 2.81×10^{-3} in the dam and 2.72×10^{-3} in the sire line.

Table 2.1: Variants detected in 70 sequenced key ancestor animals.

		Raw	*Filtered & imputed*	*Dam line*	*Sire line*	
Number of animals		70	70	32	38	
Sequence coverage[1]		16.69 (8.72 - 36.85)	18.02 (9.31 - 36.85)	15.57 (8.72 - 27.73)		
Number of variants	All	28,407,060	26,862,369	24,358,047	24,093,052	
	Biallelic SNP	22,191,375	21,209,725	19,456,000	19,232,692	
	Biallelic INDEL	4,379,470	4,339,947	3,960,976	3,928,684	
	Others[2]	1,836,215	1,312,697	941,071	931,676	
Autosomal variants	All	27,582,843	26,198,587	23,774,053	23,531,919	
	Biallelic SNP	21,553,323	20,715,354	19,015,058	18,808,294	
	Biallelic INDEL	4,248,742	4,211,012	3,846,008	3,817,622	
	Others[2]	1,780,778	1,272,221	912,987	906,003	

[1] estimated from the autosomes
[2] this category contains multi-allelic SNP, multi-allelic INDEL, as well as sites that may contain both SNP and INDEL

2.3.1 Comparison between array-called and sequence-called genotypes

Sixty-eight boars (32 and 36 from the dam and sire line, respectively) that had average sequencing coverage between 8.72 and 36.85-fold (average: 16.79-fold) also had Illumina PorcineSNP60 BeadChip-called genotypes. Using the array-called genotypes at 54,600 autosomal SNP for which we were able to determine reference and alternate alleles as a truth set, we calculated genotype concordance, non-reference sensitivity and non-reference discrepancy between array-called and sequence-called

genotypes as proposed by DePristo et al. [19]. Of the 54,600 SNP, 6,376 and 1,029 were fixed for the reference and alternate allele, respectively, and 47,195 were polymorphic in the array-called genotypes of the 68 pigs.

Of the 48,224 SNP that were either polymorphic or fixed for the alternate allele in the array-called genotypes, 46,009 (95.41%) and 45,951 (95.29%) were also present in the raw and filtered sequence variants, respectively. 1,232 SNP of the Illumina PorcineSNP60 BeadChip complement were missing in the sequenced set because they were either genotyped as INDEL or multiallelic sites using GATK and thus excluded from the comparison due to incompatible alleles. 983 and 1,041 SNP were not among the raw and filtered sequence variants, respectively, although the frequency of the minor allele was > 5% in the array-called genotypes for most (> 80%) of them. It is likely that these variants could not be matched with the sequence set due to either incompatible or ambiguous map coordinates.

Non-reference sensitivity was greater than 99% and non-reference discrepancy around 1% for the raw genotypes called by the GATK, suggesting that the high sequencing coverage facilitated accurate variant discovery (Table 2.2). The concordance between sequence- and array-called genotypes improved slightly after applying site-level hard filtration. Beagle phasing and imputation further increased the concordance and non-reference sensitivity as well as decreased the non-reference discrepancy of the filtered sequence variant genotypes.

Table 2.2: Comparison between sequence- and array-called genotypes at corresponding positions.

Dataset	Genotype concordance (%)	Non-reference sensitivity (%)	Non-reference discrepancy (%)
Raw	99.18	99.75	1.11
Filtered	99.19	99.77	1.09
Filtered & imputed	99.82	99.95	0.24

2.3.2 Population structure and genetic diversity

To investigate the population structure, ancestry and genetic diversity among the 70 sequenced pigs, we performed principal component, admixture and fixation index (F_{ST}) analyses. The principal components were extracted from a genomic relationship matrix constructed from 23,691,198 autosomal sequence variants that had minor allele frequency greater than 0.01.

The first principal component of the genomic relationship matrix explained 8.61% of the variation and separated the animals by lines (Figure 2.1B). The second principal component explaining 2.68% of the variation revealed variability within the sire line. Five outlier animals along the second axis of variation descended from imported Large White boars.

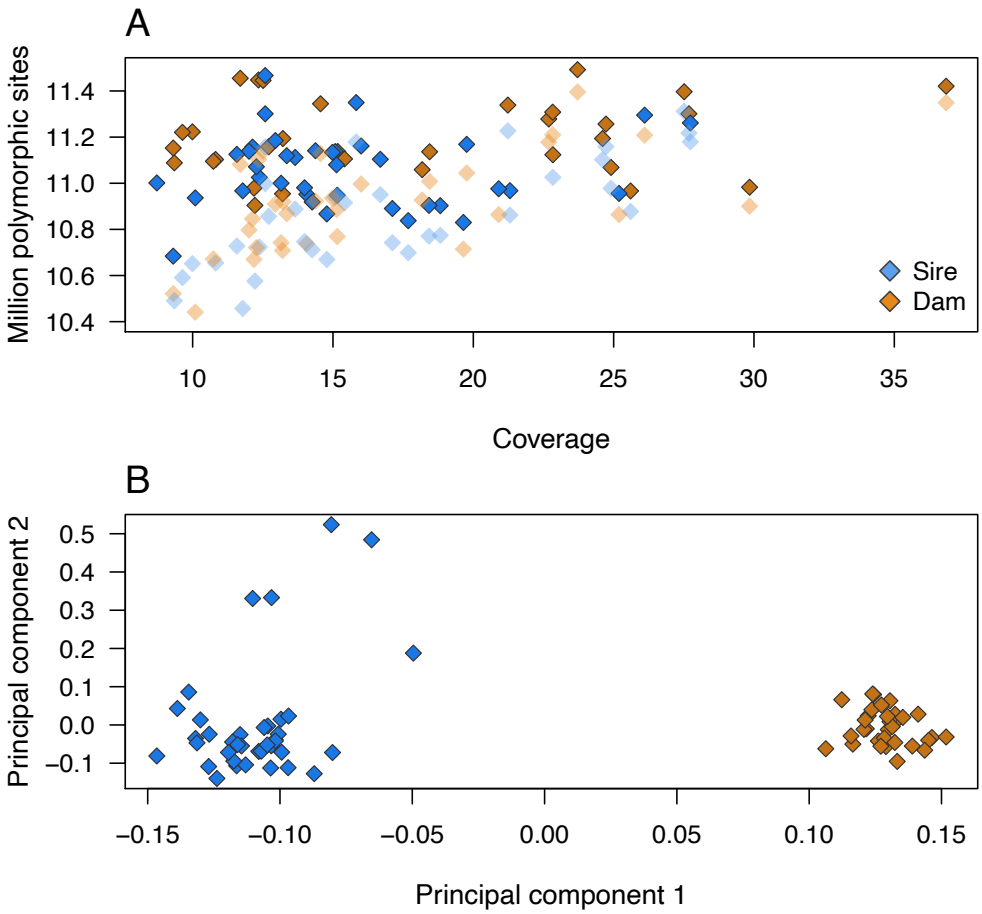


Figure 2.1: Sequencing of key ancestor animals from two pig lines.

(A) Number of polymorphic sites detected in the 70 boars as a function of depth of coverage based on imputed and filtered non-imputed data (transparency). (**B)** Plot of the first two principal components showing the separation of animals by breed and the relationship between both lines. Blue and orange symbols indicate 38 and 32 boars from the sire and dam line, respectively.

We performed an admixture analysis using 1,207,189 independent biallelic SNP to assess gene flow between both lines. As expected, K = 2 was the most plausible number of genetically distinct clusters (Supplementary Fig. S2.1). The cross-validation error for K = 1, K = 2 and K = 3 was 0.561, 0.546 and 0.564, respectively.

In order to investigate if pronounced allele frequency differences exist between both lines, we performed a SNP-based genetic differentiation analysis. We observed multiple 10 kb sliding windows scattered throughout the genome with F_{ST} values greater than 0.25, indicating genetic divergence of both lines (Supplementary Fig. S2.2, Additional file 3.2). The average weighted F_{ST} value across all windows was 0.07.

We estimated runs of homozygosity (ROH) for 19,146,365 biallelic SNP to investigate genomic inbreeding in both lines. In total 111,201 ROH with an average length of 391.28 kb (ranging from 50 kb to 11.1 Mb) were detected (Phred-scaled likelihood > 70). The ROH contained an average number of 3,176 SNP (ranging from 29 to 87,699). The boars from the dam and sire line had 1,604 ± 133 and 1,575 ± 91 ROH with an average size of 377,928 and 402,731 bp, respectively. The genomic inbreeding (F_{ROH}, i.e., the fraction of the autosomal genome covered by ROH), was 0.26 ± 0.03 and 0.28 ± 0.03 for the dam and sire line, respectively. We classified the ROH into short (50 - 100 kb), medium (100 kb - 2 Mb) and long ROH (above 2 Mb) (Figure 2.2). Most ROH belonged to the medium length class. The average F_{ROH} was similar in both lines for small and medium ROH. However, F_{ROH} was higher for long ROH in the sire line.

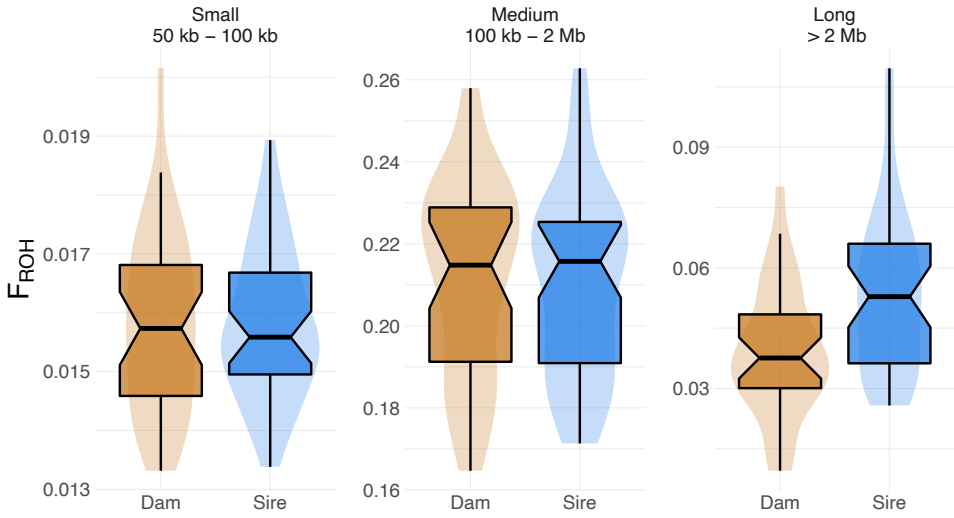


Figure 2.2: Genomic inbreeding in the two lines.

FROH in dam and sire line, estimated for three groups of ROH classified based on their length: small (50 kb – 100 kb), medium (100 kb – 2 Mb) and long (> 2 Mb).

2.3.3 Variant annotation

In 32 boars from the dam line, we annotated 23,774,053 (19,087,807 SNP; 4,038,170 INDEL) variants, including 2,567,754 variants that were not detected in the sire line. In 38 boars of the sire line, we annotated 23,531,919 (18,881,067 SNP; 4,009,043 INDEL) variants, including 2,325,620 that were not detected in the dam line. When compared to 63,832,658 germline variants listed for *Sus scrofa* in the Ensembl database (release 101), 5,745,790 (24.17%, dam line) and 5,693,068 (24.19%, sire line) variants were novel, of which the majority were INDEL and 14.66% and 14.64% were biallelic SNP.

We used the Ensembl Variant Effect Predictor software (VEP, release 98; [20]) to predict functional consequences for the sequence

variants (Table 2.3). In total, 2.96% (dam line) and 2.94% (sire line) of the variants were in exons. Putative impacts of missense variants on protein function were predicted using the SIFT (sorting intolerant from tolerant) scoring algorithm [21] as implemented in the VEP software. The scoring algorithm classified 12,024 and 11,958 amino acid substitutions in the dam and sire line, respectively, as "deleterious" (SIFT score < 0.05).

Table 2.3: Predicted consequences of variants segregating in two lines. The table shows only the most sever consequence for a variant.

Consequence type (most severe)	Dam line	Sire line
Splice donor variant	1,396	1,421
Splice acceptor variant	1,126	1,096
Stop gained	1,615	1,604
Frameshift variant	10,912	11,043
Stop lost	595	587
Start lost	423	421
Inframe insertion	990	987
Inframe deletion	1,164	1,186
Protein altering variant	62	62
Missense variant	70,758	69,983
Splice region variant	22,493	22,148
Incomplete terminal codon variant	12	11
Synonymous variant	76,977	75,279
Stop retained variant	149	135
Start retained variant	4	4
Coding sequence variant	98	96
Mature miRNA variant	12	16
5' - UTR variant	168,000	164,866
3' - UTR variant	348,135	344,514
Non-coding transcript exon variant	277,002	275,909
Intron variant	12,213,614	12,092,056
Non-coding transcript variant	11	10
Upstream gene variant	878,779	869,207
Downstream gene variant	757,364	750,548
Intergenic variant	8,942,362	8,848,730

2.3.4 Known trait-associated variants

The catalogue of Mendelian traits in *Sus scrofa* curated in the OMIA database (https://omia.org/home/, [22]) contained records of 47 likely causal variants (as of September 2020). However, the genomic coordinates were available for only 33 likely causal variants. Using functional annotations and sequence coverage analyses, we detected OMIA-listed variants affecting the *KIT*, *MC1R* and *FUT1* genes in the sequenced key ancestor animals that occurred at alternate allele frequencies between 0.013 and 1 (Supplementary Table S2.1, Additional file 3.3).

A duplication of the *KIT* gene and a splice site variant in intron 17 of the *KIT* gene are associated with the dominant white phenotype [23, 24]. Because the genotyping of larger structural and copy number variants from short-read sequencing data is notoriously difficult, we visually inspected the depth of sequencing coverage at the SSC8 region encompassing *KIT*. An increase in coverage between 41.22 and 41.78 Mb confirmed the presence of a previously reported 560 kb duplication (DUP1; Supplementary Fig. S2.3, Additional file 3.4, [25, 26]. The duplication also encompasses a copy of *KIT* that carries a splice donor site variant (SSC8: 41486012G>A, rs345599765) which manifests in a dominant white phenotype [23, 24]. The splice variant segregated at a frequency of 0.49 and 0.42 in the sire and dam line, respectively. Seven animals that carried either one or two copies of DUP1 did not carry the splice site variant and all others were heterozygous carriers. Because this variant is located within the 560 kb duplication, we observed allelic imbalance in heterozygous animals.

We detected three OMIA-listed pigmentation-associated variants in the *MC1R* gene in the sequenced pigs. All boars were homozygous carriers

of a 2-bp insertion (SSC6: 182,120 - 182,121 bp), that causes a frameshift and premature translation termination, which is associated with recessive white color [27]. All animals were also homozygous carriers of two missense variants in the *MC1R* gene (SSC6: 181461T>C, ENSSSCP00000027395.1: p.Thr243Ala and SSC6: 181697A>G, ENSSSCP00000027395.1: p.Val164Ala), for which the reference alleles had been associated with red color in the Duroc breed [28].

A missense variant (SSC6: 54079560T>C; ENSSSCP00000062180.1: p.Thr102Ala; rs335979375) in the *FUT1* gene enables adhesion of enterotoxigenic Escherichia Coli F18 fimbriae (ETEC F18) to receptors at the brush border membranes of the intestinal mucosa [29]. The allele that facilitates ETEC F18 adhesion causes diarrhea in neonatal and recently weaned piglets. Since a strong selection against the ETEC F18 susceptible allele takes place in both SLW lines, we observed the disease-associated allele only in one boar from the sire line in the heterozygous state.

2.3.5 Signatures of selection

We detected signatures of past selection using the composite likelihood ratio (CLR) test. Signatures of ongoing selection were identified by the integrated haplotype score (iHS) test. For both analyses, we used biallelic autosomal SNP (N_{dam} = 19,015,058, N_{sire} = 18,808,294) that were grouped into non-overlapping 100 kb windows. For the CLR tests, we considered an empirical 0.5% significance threshold to identify putative signatures of selection (Figure 2.3A). The number and length of candidate selection regions was higher in the dam than the sire line (14 vs. 7; 38.1 Mb vs. 26.1 Mb). Two regions on SSC3 (from 122.6 to 124.9 Mb) and SSC13 (from 140.0 to 146.1 Mb) showed evidence of selection in both lines. For the iHS analyses, we used an empirical 0.1% significance

threshold to detect putative signatures of selection (Figure 2.3B). We detected 14 and 16 candidate regions of selection in the dam and sire line, respectively, encompassing 28.5 Mb and 32.5 Mb. Four regions on SSC1 (from 51.1 to 53.7 Mb, from 142.7 to 146.2 Mb), SSC6 (from 64.9 to 69.3 Mb) and SSC13 (from 148.0 to 150.6 Mb) were shared between both lines.

Considering both statistics, we detected more signatures of selection in the dam than sire line (28 vs. 23). Only 6 regions, detected by either CLR or iHS, overlapped between both lines. A strong signature of selection was detected in both lines with both methods on SSC13 between 140 and 152.4 Mb. The candidate region encompassed 125 genes (Supplementary Table S2.2, Additional file 3.5), as well as 63,480 and 55,835 polymorphic sites in the dam and sire line, respectively, precluding to readily prioritize candidate genes and variants responsible for the sweep.

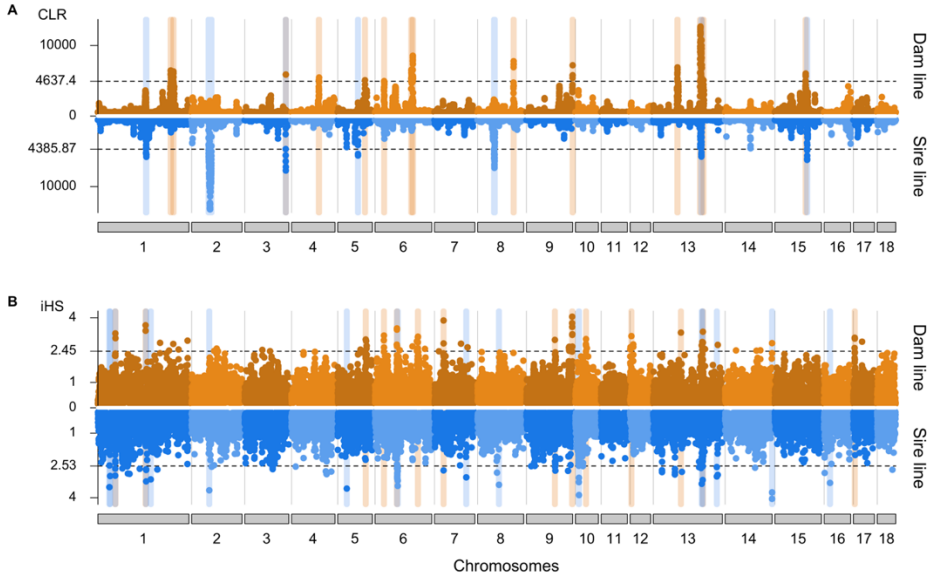


Figure 2.3: Signatures of selection detected in the sire and dam line of the SLW breed.

Signatures of selection detected in the sire and dam line of the SLW breed using CLR **(A)** and iHS **(B)** Dotted lines indicate the empirical 0.5 (CLR) and 0.1%

(iHS) thresholds. Blue, orange and grey vertical bars highlight signatures of selection detected in the sire, dam and both lines, respectively.

2.3.6 Reference-based genotyping from low-coverage sequencing data

In order to investigate if the 70 sequenced key ancestor animals may serve as a reference panel for genotyping by low-coverage sequencing, we sequenced the genomes of 175 pigs (84 from the sire line and 91 from the dam line) at low coverage using Gencove's low-pass sequencing solution. The pigs also had Illumina PorcineSNP60 BeadChip-called genotypes. A principal component (Supplementary Fig. S2.4, Additional file 3.6) analysis of a genomic relationship matrix constructed from microarray-derived genotypes showed that the 175 pigs cluster with the 70 key ancestor animals.

Following quality control, we aligned a median number of 16,153,314 (between 5,950,534 and 21,168,683) read pairs (2 x 150 bp) to the porcine reference genome, achieving an average depth of coverage of 1.11-fold (from 0.38 to 1.51). On average, 54% of the reference nucleotides were covered with at least one read. Following the reference-guided low-pass sequence variant genotyping approach (GLIMPSE) proposed by Rubinacci et al. [8], we utilized the haplotypes of the 70 sequenced key ancestor animals as a reference panel to call genotypes at 22,618,811 polymorphic sites in the 175 low-pass sequenced samples.

We assessed the accuracy of genotyping by low-pass sequencing based on Illumina PorcineSNP60 BeadChip-called genotypes at 54,600 SNP, for which we were able to determine reference and alternate alleles. Of the 54,600 SNP, 6,176 and 965 were fixed for the reference and alternate allele, respectively, in the 175 pigs according to the array-called genotypes.

Of 48,424 SNP that were either polymorphic or fixed for the alternate allele, 46,001 (94.99%) were also among the GLIMPSE-imputed genotypes. 2,423 SNP had microarray-derived genotypes but were missing in the GLIMPSE-imputed genotypes because these SNP were missing in the haplotype reference panel constructed from the key ancestor animals.

The genotype concordance, non-reference sensitivity and non-reference discrepancy between GLIMPSE-imputed and array-called genotypes at 46,001 autosomal SNP was 97.60, 98.73 and 3.24% in 175 low-pass sequenced pigs (Table 2.4, Figure 2.4A). When the sequence variant calling of the 175 samples was performed together with the 70 key ancestor animals using the multi-sample approach implemented in the GATK, all concordance metrics were considerably worse. Although, Beagle imputation improved the genotype calls of GATK for the low-pass sequenced samples, the genotype concordance and non-reference sensitivity was lower and non-reference discrepancy higher using GATK than GLIMPSE. Using the GLIMPSE approach improved the genotype concordance over GATK filtered & Beagle imputed variants by 13.83% and this improvement is mostly due to a lower non-reference discrepancy (Table 2.4).

Table 2.4: Accuracy of sequence variant genotyping in low-coverage (1.11-fold) sequencing data.

Variant genotyping approach	Genotype concordance	Non-reference sensitivity	Non-reference discrepancy
GLIMPSE	97.60	98.73	3.24
GATK raw	75.90	52.35	30.20
GATK filtered	75.89	52.36	30.22
GATK filtered & imputed	85.74	96.56	19.34

We constructed genomic relationship matrices (GRM) from the microarray-derived and GLIMPSE-imputed genotypes of the 175 sequenced pigs based on a subset of 44,268 SNP that were detected at minor allele frequency greater than 0.01 in both datasets. Both the off-diagonal and the diagonal elements of the GRM constructed from array-derived genotypes had greater variance ($\sigma^2_{diag} = 3.37 \times 10^{-3}$, $\sigma^2_{off} = 9.34 \times 10^{-3}$) than corresponding elements of the GRM constructed from low-pass sequencing data ($\sigma^2_{diag} = 3.30 \times 10^{-3}$, $\sigma^2_{off} = 9.15 \times 10^{-3}$). While the correlation of the off-diagonal (r = 0.99) and diagonal (r = 0.96) elements was high between both GRMs, the values of the diagonal elements were higher for all samples using the GLIMPSE-imputed than microarray-derived genotypes (Figure 2.4B and 4C). The average value of the diagonal elements of the GRM was 1.01 ± 0.06 and 1.05 ± 0.06 for the microarray- and low-pass sequencing-derived genotypes, respectively. On average, the 175 boars were homozygous for 65.58 ± 1.39% and 67.27 ± 1.49% of the 44,268 SNP when the genotypes were called from the microarray and low-pass sequencing data, respectively.

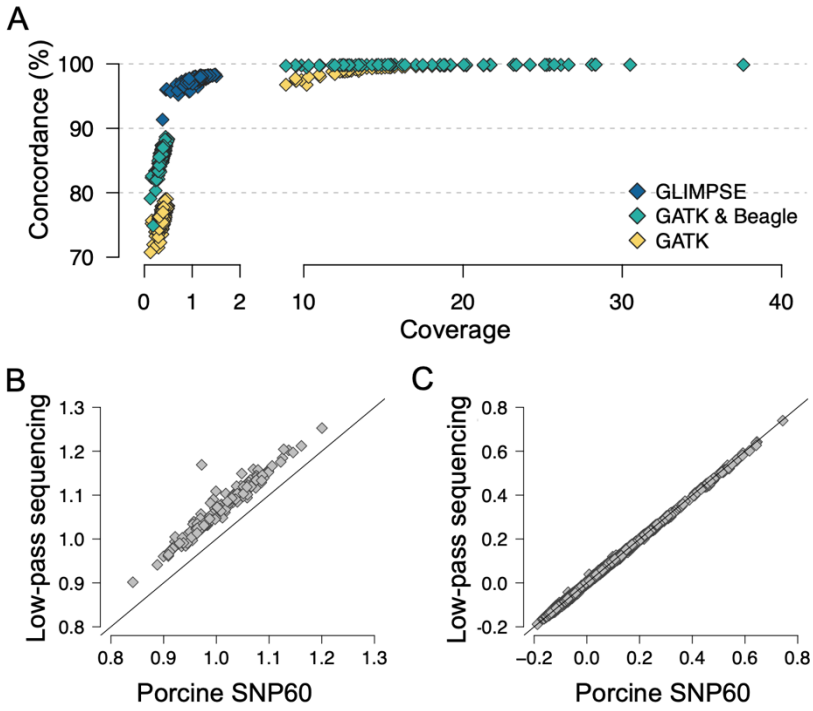


Figure 2.4: Accuracy of genotyping by low-coverage sequencing.

(A) Concordance between array-called and sequence-called genotypes at 46,001 biallelic autosomal SNP in 243 pigs that had been sequenced at either low (N = 175; < 1.5-fold) or medium to high (N = 68; 8.88 - 37.60-fold) coverage. Correlations of (**B)** diagonal (r = 0.96) and (**C)** off-diagonal (r = 0.99) elements of genomic relationship matrices constructed from array- and GLIMPSE-called genotypes at 44,268 SNP that had minor allele frequency > 0.01.

2.4 Discussion

We applied a key ancestor animal approach to prioritize 38 and 32 boars that accounted for 95.35 and 87.95% of the genetic diversity of the SLW sire and dam line, respectively. The contributions of the SLW key ancestor animals to the current populations are considerably higher than reported for other populations. For instance, 43 key ancestor animals

explained 69% of the genetic diversity of the Fleckvieh cattle population [5]. Neuditschko et al. [30] selected 41 and 55 key contributors, respectively, that explained 78% and 75% of the genetic relationship structure of the Swiss Franches-Montagnes horse and Australian Holstein-Friesian cattle population. The effective population size of the SLW sire and dam line is 44 and 72, respectively, which is less than half the effective population size of the Fleckvieh cattle and Swiss Franches-Montagnes horse population [31, 32]. Thus, a few animals that are selected based on their marginal genetic contribution to the active breeding population, account for a large fraction of the population's haplotype diversity. It is worth mentioning that approaches other than the key ancestor animal approach may increase the haplotype diversity among the sequenced animals [33]. Nevertheless, the catalogue of 26.86 million polymorphic sites detected from the 70 sequenced boars of our study contains most alleles that segregate in the SLW populations, particularly those that occur at not too low frequency. A Ti/Tv-ratio of 2.28 indicates that the variants were of high quality [34]. In spite of the low effective population size, the nucleotide diversity (π) was high in both lines (π_{dam} = 2.24 x 10^{-3}; π_{sire} = 2.23 x 10^{-3}), which agrees well with estimates obtained in other European pig populations [35–37]. The nucleotide diversity in the SLW populations is higher than in cattle (1.77 x 10^{-3} – 1.90 x 10^{-3}) and human (0.98 x 10^{-3} – 1.41 x 10^{-3}) populations that have considerably larger current effective population sizes [38].

Although our sequencing cohort contained more animals from the sire line, we detected somewhat more autosomal variants in the dam line (N_{sire} = 23,531,919; N_{dam} = 23,774,053). While the average number of heterozygous variants detected per animal was higher in the dam line (N_{sire} = 6,180,048; N_{dam} = 6,351,565), the number of variants homozygous for

the alternate allele was higher in the sire line (N_{sire} = 4,873,994; N_{dam} = 4,646,236). Because the average depth of sequencing was similar in the sire and dam line, these differences are unlikely to be due to uneven coverage between the lines. These differences are likely attributable to a smaller effective population size and higher genomic inbreeding in the sire line. The presence of many long ROH (> 2 Mb) suggests that recent inbreeding is higher in the sire than the dam line. Small effective population size and increasing inbreeding make both lines susceptible to the phenotypic manifestation of recessive alleles. For instance, a recessive sperm defect has recently been discovered in the sire line [39]. The management of an ever-increasing number of recessive traits is a challenge to domestic animal breeding populations [40–42]. Efficient and sustainable strategies are required to prevent the frequent manifestation of recessive diseases in populations with low effective population size.

Surprisingly, Cai et al. [43] detected fewer variants (between 20.68 and 22.11 million variants) in a considerably larger cohort of pigs (between 61 and 89) from three commercial Danish lines. Considering that Cai et al. also sequenced key ancestor animals, this difference to our study suggests higher genetic diversity in SLW. However, the depth of coverage, sequencing strategy, sequence variant genotyping and filtration approaches have major impacts on detecting polymorphic sites [44, 45]. While the effective depth of coverage realized by Cai et al. [43] is unknown to us, our samples were sequenced at an average depth of coverage greater than 16-fold. This depth of coverage enabled us to accurately detect both homozygous and heterozygous sites as evidenced by high non-reference sensitivity and genotype concordance at low non-reference discrepancy.

The principal components of a genomic relationship matrix constructed from whole-genome sequence variants revealed a separation of the animals by line. While the differentiation between the two populations might be less evident if diverse samples or an outgroup were considered in the analysis [7, 37, 46, 47], an average F_{ST} value of 0.07 corroborated that both lines diverged considerably. In fact, the average F_{ST} value observed between two SLW lines is similar to values reported between distinct European pig breeds [37, 43, 48]. The differentiation between the sire and dam line might result from distinct breeding objectives with negative genetic correlations [49]. While the sire line is mainly selected for meat and fattening traits, the dam line is mainly selected for reproduction traits. Using CLR and iHS, we detected 51 candidate signatures of selection, of which only six overlapped between both lines, suggesting that different loci are under selection in the sire and dam line. However, previous research indicates that selection for complex traits, such as production and reproduction, acts on many loci, thus barely leaves strong footprints in the genome [50, 51]. Moreover, both lines diverged only few generations ago, rendering limited time for shifts in allele frequency due to selection. We suspect that the strong differentiation between the SLW sire and dam line is also a result of genetic drift [52–54] due to very small effective population size and pronounced founder effects resulting from the unbalanced use of individual boars in artificial insemination.

A reference panel of less than 70 sequenced key ancestor animals facilitated imputing sequence variant genotypes at high accuracy and detecting trait-associated nucleotides using genome-wide association testing in cattle populations [6, 55]. Sequence variant genotypes are typically inferred using two-step imputation approaches. This requires the presence of a representative number of animals that had been genotyped at

high density [11]. However, routine genotyping in the SLW populations is performed using a customized PorcineSNP60 BeadChip. Genotypes from high-density microarrays (e.g., 600K) are not available. Thus, precluding the accurate imputation of sequence variant genotypes from the key ancestor animals using the well-established stepwise imputation approach [13]. This limitation prompted us to investigate an alternative approach to reference-guided sequence variant imputation. We considered the 70 key ancestor animals as a reference to call genotypes from low-pass sequencing data (1.11-fold) of genetically similar pigs. In agreement with previous studies in human and cattle populations, the genotyping accuracy from the low-pass sequencing data was very high [8, 15, 56]. Moreover, the low-pass sequencing-derived genomic relationship coefficients were highly correlated with those obtained using microarray genotyping. This suggests that the low-pass sequencing-derived imputed genotypes may readily be used for genomic prediction [56, 57]. However, the diagonal elements of the genomic relationship matrix were higher and had less variance using the genotypes from low-pass sequencing than microarray genotyping, likely because the sequenced key ancestor animals do not represent the full haplotype diversity of the SLW populations which precludes the imputation of rarer sites that predominantly occur in the heterozygous state. High-coverage sequencing of few additional animals that carry rare haplotypes may mitigate this ascertainment bias [49] and increase the accuracy of genotyping by low-pass sequencing, particularly for rare alleles. While a subset of the 22.62 million variants obtained is sufficient to accurately predict genomic breeding values, the full variant catalogue, once available for a large mapping cohort, will facilitate powerful genome-wide association studies at nucleotide resolution.

2.5 Conclusions

The high-coverage sequencing of 70 key ancestor animals from two SLW lines and subsequent reference-guided variant discovery revealed 26,862,369 polymorphic sites. Population-genetic analyses suggest considerable genetic differentiation between both lines. Our results indicate that the key ancestor genomes may serve as a haplotype reference panel for genotyping by low-pass sequencing at high accuracy in the Swiss pig breeds. Using genotyping by low-pass sequencing increases the variant density over the currently used microarray by > 350-fold, thus providing a valuable resource for powerful genome-wide association testing.

2.6 Methods

2.6.1 Animals and whole-genome sequencing

Whole genome sequence data were generated for 70 boars. Sixty-five boars (32 from the dam line and 33 from the sire line) were selected based on their marginal genetic contribution to the current breeding populations using a key ancestor approach [3, 4]. The marginal genetic contribution was estimated based on a numerator relationship matrix that was constructed using the PyPedal python package [59]. The effective population size of the sire and dam line was estimated based on the difference in pedigree-derived inbreeding coefficients between active breeding animals and their parents following equation (3) presented in Leroy et al. [60]. The inbreeding coefficients were extracted from the numerator relationship matrix. Animals born after 01.01.2018 were considered as active breeding animals. In addition, we considered whole-genome sequence data from five boars from the sire line that were generated previously [39]. DNA was prepared from preserved blood samples that were provided by SUISAG

(the Swiss competence center for pig breeding). No animals were specially sampled for the present study. Illumina TruSeq PCR-free libraries with insert sizes of 350 bp were prepared and sequenced with an Illumina NovaSeq6000 instrument using 2 ×150 bp paired-end reads.

2.6.2 Alignment quality, read mapping and depth of coverage

We used the fastp software [61] to remove adapter sequences and reads that had Phred-scaled quality less than 15 for more than 15% of the bases. Subsequently, the filtered reads were aligned to the SSC11.1 assembly of the porcine genome [62] using the mem-algorithm of the BWA software [63]. The Picard tools software suite [64] and Sambamba [65] were applied to mark duplicate reads and sort the alignments by coordinates, respectively. To calculate depth of coverage, we extracted the number of reads covering a genomic position using the mosdepth software [66]. For the coverage calculation, we discarded reads with mapping quality < 10 and SAM bitwise flag value of 1796.

2.6.3 Variant calling

We used the BaseRecalibrator module of the Genome Analysis Toolkit (GATK - version 4.1.0; [19]) to adjust the base quality scores while supplying 63,881,592 unique positions from the porcine dbSNP version 150 as known variants. We applied the HaplotypeCaller, GenomicsDBImport and GenotypeGVCFs modules from the GATK to discover and genotype SNP and INDEL in the 70 SLW pigs together with 28 samples from various breeds that were sequenced earlier. Subsequently, we applied the VariantFiltration module of the GATK according to best practice recommendations for site-level hard filtration to retain high-

quality variants. Beagle (version 4.1; [18]) haplotype phasing and imputation was applied to impute sporadically missing sites and improve the primary genotypes obtained using the GATK.

The concordance between sequence- and array called genotypes was calculated for 68 pigs that also had Illumina PorcineSNP60 BeadChip microarray-derived genotypes. We considered only autosomal SNP. We converted the TOP/BOT alleles of the microarray-derived genotypes to REF/ALT allele coding to make them compatible with the sequence-derived genotypes. This was possible for 54,600 SNP. Sequence variant genotyping accuracy was quantified using genotypic concordance, non-reference sensitivity and non-reference discrepancy [19, 44].

Invariant sites and variants within regions with an excessive depth of coverage (> mean coverage + 2 * SD) were removed using VCFtools (v. 0.1.16; [67]). The resulting data were split into two datasets containing 23,774,053 and 23,531,919 variants segregating in 32 boars from the dam line and 38 boars from the sire line, respectively.

2.6.4 Functional annotation

Functional consequences of the variants (including SIFT scores [21] for missense variants) were predicted with the Ensembl Variant Effect Predictor (VEP, version 91.3; [20]) using local cache files from Ensembl release 98. The transition to transversion ratio (Ti/Tv) was calculated using BCFtools command *stats* (version 1.8; [68]).

2.6.5 Detection of mendelian trait-associated variants and coverage analysis

We downloaded genomic coordinates of 47 likely causal variants from the Online Mendelian Inheritance in Animals (OMIA) database [22]. Genes harboring likely causal variants for which the genomic coordinates

were not annotated according to SSC11.1 were manually inspected. Read alignments and sequence coverage in regions harboring known larger structural variants were manually inspected.

2.6.6 Population structure and genetic diversity analysis

The structure of the two lines was investigated using ADMIXTURE (v1.3.0; [69]). To avoid confounding due to extensive linkage disequilibrium (LD), we removed correlated loci based on high levels ($r^2 > 0.6$) of pairwise LD using PLINK (version 1.9; [70]) with the "--indep-pairwise 100 25 0.6" option before running the ADMIXTURE analysis. The number of ancestral clusters (K) was set from 1 to 3, and five-fold cross-validation was performed to determine the K value with the lowest cross-validation error.

A genomic relationship matrix was built using 23,691,198 autosomal sequence variants that had a minor allele frequency higher than 0.01 using PLINK. The principal components of the genomic relationship matrix were calculated using the GCTA (version 1.92.1; [71]) software. We applied the GCTA flag "--grm-singleton" to identify four pairs of animals with relationship coefficients ranging from 0.32 to 0.37. One animal from each pair was removed for the F_{ST} and signature of selection analyses (1 from the dam line and 3 from the sire line).

We calculated the weighted genome wide fixation index (F_{ST}, [72]) based on pairwise differences in the variances of allele frequencies using 24,926,366 biallelic variants. F_{ST} values were calculated in 10 kb sliding windows with an overlap of 5 kb using the "--weir-fst-pop" flag of VCFtools (v.1.2.11; [67]). The manhattan plot was constructed using the R package qqman [73].

Nucleotide diversity (π) was calculated over all biallelic autosomal variants in 10 kb sliding windows with an overlap of 5 kb using VCFtools.

Runs of homozygosity (ROH) were estimated with BCFtools/ROH [74] using the GATK-derived genotypes (containing the Phred-scaled likelihoods). We considered biallelic SNP that had non-missing genotypes in all animals (maximal missing count per site was set to 0). According to Tortereau et al. [75], we assumed a constant recombination rate of 0.7 cM/Mb along the chromosomes. Average genomic inbreeding (F_{ROH}) was calculated assuming an autosomal genome length of 2,265,774,640 bases. Following a recent study by Bhati et al. [76], we classified the ROH based on their length (short: 50 - 100 kb, medium: 100 kb - 2 Mb, long: > 2 Mb).

2.6.7 Signatures of selection (CLR and iHS) and candidate regions

Putative signatures of selection were detected using integrated Haplotype Scores (iHS) and composite likelihood ratios (CLR). The iHS [77] reveals 'soft sweeps', i.e., signatures of selection where selection for beneficial alleles is still ongoing. The CLR [78] reveals 'hard sweeps', i.e., signatures of selection where beneficial alleles recently reached fixation. We considered 24,926,366 autosomal biallelic SNP from 31 and 35 boars from the dam and sire line, respectively. The genotypes were phased using Beagle (version 5.1; [79]) with disabled imputation and effective population size set to 50. The CLR statistic was calculated chromosome-wise with the SweepFinder2 software [80] using a pre-computed empirical allele frequency spectrum and 100 kb spacing between test sites (-lg 100000). Using the R package rehh 2.0 [81], we applied the function *scan_hh* to estimate the integrated extended haplotype homozygosity (EHH) on variants with MAF > 0.05 for each chromosome

separately. Subsequently, we applied the function *ihh2ihs* to obtain standardized iHS values in 100 kb non-overlapping windows.

The function *calc_candidate_regions* from the rehh 2.0 package [81] was applied to select candidate signatures of selection in 100 kb windows using the parameters "window_size = 1E6", "overlap = 1E5", "pval = F" and "min_n_extr_mrk = 1". Empirical significance thresholds were chosen after visual inspection of the distribution of the test statistics (0.1% in iHS and 0.5% in CLR). Genes overlapping with candidate signatures of selections were determined based on the Ensembl (release 98) annotation of the porcine genome.

2.6.8 Analysis of low-pass sequence data

A median number of 16,131,419 paired-end (2x150bp) reads were generated for 96 pigs from the dam line and 96 pigs from the sire line. Adapter sequences and bases and reads with low sequencing quality were removed with fastp [61]. Subsequently, the reads were aligned to the porcine reference genome (SSC11.1) using the mem-algorithm of BWA [63] and duplicate reads were marked using Samblaster [82]. Following the read alignment, six samples were excluded because the mapping rate and the proportion of properly paired reads was less than 70 and 75%, respectively. Additionally, we excluded 10 samples for which the average coverage was less than 0.2-fold and one sample for which ancestry could not be verified.

To compile the reference haplotypes, we retained 22,618,811 biallelic autosomal SNP that were polymorphic (minor allele count ≥ 1) among the 70 key ancestor pigs. Following the approach proposed by Rubinacci et al. [8], we used the *mpileup* and *call* commands of BCFtools

[68] to calculate genotype likelihoods at the 22,618,811 polymorphic sites in the 175 low-pass sequenced and reference-aligned samples. Subsequently, we applied the phasing and imputation algorithm implemented in GLIMPSE_phase [8] to refine the BCFtools-derived genotype calls using the previously established haplotype reference panel. This approach produced genotypes at 22,618,811 sites for the 175 low-pass sequenced samples. A genomic relationship matrix among the low-pass sequenced animals was constructed from the low-pass sequencing data-derived genotypes using GCTA [71].

2.7 List of abbreviations

CLR: Composite likelihood ratio; F_{ST}: Fixation index; iHS: Integrated haplotype score; INDEL: Insertions and deletions; LD: Linkage disequilibrium; MAF: Minor allele frequency; OMIA: Online Mendelian Inheritance in Animals; PCA: Principal component analysis; QTL: Quantitative trait loci; ROH: Run of homozygosity; SLW: Swiss Large White; SNP: Single nucleotide polymorphism; Ti/Tv: Transition/transversion ratio; WGS: Whole-genome sequencing

2.8 References

1. SUISAG. https://www.suisag.ch/. Accessed 5 Apr 2021.
2. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE. 2009;4:e6524–e6524.
3. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for the allocation of sequencing resources in genotyped livestock populations. Genetics Selection Evolution. 2017;49:1–16. doi:10.1186/s12711-017-0322-5.

4. Goddard ME, Hayes BJ. Genomic Selection Based on Dense Genotypes Inferred From Sparse Genotypes. Proc Assoc Advmt Anim Breed Genet. 2009;18.

5. Jansen S, Aigner B, Pausch H, Wysocki M, Eck S, Benet-Pagès A, et al. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics. 2013;14:446.

6. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics. 2014;46:858–65.

7. Bovo S, Ribani A, Muñoz M, Alves E, Araujo JP, Bozzi R, et al. Whole-genome sequencing of European autochthonous and commercial pig breeds allows the detection of signatures of selection for adaptation of genetic resources to different breeding and production systems. Genetics Selection Evolution. 2020;52:1–19. doi:10.1186/s12711-020-00553-7.

8. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nature Genetics. 2021;53:120–6. doi:10.1038/s41588-020-00756-0.

9. Gusev A, Shah MJ, Kenny EE, Ramachandran A, Lowe JK, Salit J, et al. Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. Genetics. 2012;190:679–89.

10. Zan Y, Payen T, Lillie M, Honaker CF, Siegel PB, Carlborg Ö. Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: A cost-efficient approach. Genetics Selection Evolution. 2019;51:1–11. doi:10.1186/s12711-019-0487-1.

11. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genetics Selection Evolution. 2017;49:24.

12. van den Berg S, Vandenplas J, van Eeuwijk FA, Bouwman AC, Lopes MS, Veerkamp RF. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. Genetics Selection Evolution. 2019;51:1–13. doi:10.1186/s12711-019-0445-y.

13. van Binsbergen R, Bink MCAM, Calus MPL, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution. 2014;46:41. doi:10.1186/1297-9686-46-41.

14. Ros-Freixedes R, Whalen A, Gorjanc G, Mileham AJ, Hickey JM. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. Genetics Selection Evolution. 2020;52.

15. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Research. 2021;:gr.266486.120. doi:10.1101/gr.266486.120.

16. Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, et al. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. BMC Genomics. 2021;22:197. doi:10.1186/s12864-021-07508-2.

17. van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013;43 SUPL.43. doi:10.1002/0471250953.bi1110s43.

18. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics. 2007;81:1084–97.

19. Depristo MA, Banks E, Poplin R, Garimella K v., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43:491–501.

20. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17.

21. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Research. 2012;40:W452-7.

22. Nicholas FW. Online Mendelian Inheritance in Animals (OMIA): a record of advances in animal genetics, freely available on the Internet for 25 years. Animal Genetics. 2021;52:3–9. doi:10.1111/age.13010.

23. Sun G, Liang X, Qin K, Qin Y, Shi X, Cong P, et al. Functional Analysis of KIT Gene Structural Mutations Causing the Porcine

Dominant White Phenotype Using Genome Edited Mouse Models. Frontiers in Genetics. 2020;11:138. doi:10.3389/fgene.2020.00138.

24. Marklund S, Kijas J, Rodriguez-Martinez H, Ronnstrand L, Funa K, Moller M, et al. Molecular basis for the dominant white phenotype in the domestic pig. Genome Research. 1998;8:826–33.

25. Rubin C-JJ, Megens H-JJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences of the United States of America. 2012;109:19529–36. doi:10.1073/pnas.1217149109.

26. Wu Z, Deng Z, Huang M, Hou Y, Zhang H, Chen H, et al. Whole-Genome Resequencing Identifies KIT New Alleles That Affect Coat Color Phenotypes in Pigs. Frontiers in Genetics. 2019;10 MAR:218. doi:10.3389/fgene.2019.00218.

27. Jia Q, Cao C, Tang H, Zhang Y, Zheng Q, Wang X, et al. A 2-bp insertion (c.67_68insCC) in MC1R causes recessive white coat color in Bama miniature pigs. Journal of Genetics and Genomics. 2017;44:215–7.

28. Kijas JMH, Wales R, Törnsten A, Chardon P, Moller M, Andersson L. Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. Genetics. 1998;150:1177–85.

29. Meijerink E, Neuenschwander S, Fries R, Dinter A, Bertschinger HU, Stranzinger G, et al. A DNA polymorphism influencing a(1,2)fucosyltransferase activity of the pig FUT1 enzyme determines susceptibility of small intestinal epithelium to Escherichia coli F18 adhesion. Immunogenetics. 2000;52:129–36.

30. Neuditschko M, Raadsma HW, Khatkar MS, Jonas E, Steinig EJ, Flury C, et al. Identification of key contributors in complex population structures. PLoS ONE. 2017;12. doi:10.1371/journal.pone.0177638.

31. Poncet PA, Pfister W, Muntwyler J, Glowatzki-Mullis ML, Gaillard C. Analysis of pedigree and conformation data to explain genetic variability of the horse breed Franches-Montagnes. Journal of Animal Breeding and Genetics. 2006;123:114–21. doi:10.1111/j.1439-0388.2006.00569.x.

32. Pausch H, Aigner B, Emmerling R, Edel C, Götz KU, Fries R. Imputation of high-density genotypes in the Fleckvieh cattle population. Genetics Selection Evolution. 2013;45:3. doi:10.1186/1297-9686-45-3.

33. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for the allocation of sequencing resources in genotyped livestock populations. Genetics Selection Evolution. 2017;49.

34. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015;31:318–23. doi:10.1093/bioinformatics/btu668.

35. Tong X, Hou L, He W, Mei C, Huang B, Zhang C, et al. Whole genome sequence analysis reveals genetic structure and X-chromosome haplotype structure in indigenous Chinese pigs. Scientific Reports. 2020;10:1–10. doi:10.1038/s41598-020-66061-2.

36. Bosse M, Megens H-JJ, Madsen O, Paudel Y, Frantz LAFF, Schook LB, et al. Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. PLoS Genetics. 2012;8:e1003100. doi:10.1371/journal.pgen.1003100.

37. Zhang C, Plastow G. Genomic Diversity in Pig (Sus scrofa) and its Comparison with Human and other Livestock. Current Genomics. 2011;12:138–46.

38. Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. 2019;21:1–27. doi:10.1101/2019.12.20.882423.

39. Nosková A, Hiltpold M, Janett F, Echtermann T, Fang Z-H, Sidler X, et al. Infertility due to defective sperm flagella caused by an intronic deletion in *DNAH17* that perturbs splicing. Genetics. 2020. doi:10.1093/genetics/iyaa033.

40. Cole JB. A simple strategy for managing many recessive disorders in a dairy cattle breeding program. Genetics Selection Evolution. 2015;47:94. doi:10.1186/s12711-015-0174-9.

41. Derks MFL, Megens HJ, Bosse M, Lopes MS, Harlizius B, Groenen MAM. A systematic survey to identify lethal recessive variation in highly managed pig populations. BMC Genomics. 2017;18:1–12.

42. Pausch H, Schwarzenbacher H, Burgstaller J, Flisikowski K, Wurmser C, Jansen S, et al. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. BMC Genomics. 2015;16:312.

43. Cai Z, Sarup P, Ostersen T, Nielsen B, Fredholm M, Karlskov-mortensen P, et al. Animal Genetics and Genomics Genomic diversity revealed by whole-genome sequencing in three Danish commercial pig breeds. 2020;98:1–12.

44. Crysnanto D, Wurmser C, Pausch H. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. Genetics Selection Evolution. 2019;51:1–15. doi:10.1186/s12711-019-0462-x.

45. Taylor JF, Whitacre LK, Hoff JL, Tizioto PC, Kim J, Decker JE, et al. Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals. Genetics Selection Evolution. 2016;48:59. doi:10.1186/s12711-016-0237-6.

46. Ramos-Onsins SE, Burgos-Paz W, Manunza A, Amills M. Mining the pig genome to investigate the domestication process. Heredity. 2014;113:471–84. doi:10.1038/hdy.2014.68.

47. Zanella R, Peixoto JO, Cardoso FF, Cardoso LL, Biegelmeyer P, Cantão ME, et al. Genetic diversity analysis of two commercial breeds of pigs using genomic and pedigree data. Genet Sel Evol. 2016;48:24. doi:10.1186/s12711-016-0203-3.

48. Yang J, Li W-R, Lv F-H, He S-G, Tian S-L, Peng W-F, et al. Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. Molecular Biology and Evolution. 2016;33:2576–92. doi:10.1093/molbev/msw129.

49. Holm B, Bakken M, Klemetsdal G, Vangen O. Genetic correlations between reproduction and production traits in swine. Journal of Animal Science. 2004;82:3458–64. doi:10.2527/2004.82123458x.

50. Johansson AM, Pettersson ME, Siegel PB, Carlborg Ö. Genome-wide effects of long-term divergent selection. PLoS Genetics. 2010;6.

51. Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics. 2014;15:246.

52. Feder AF, Kryazhimskiy S, Plotkin JB. Identifying signatures of selection in genetic time series. Genetics. 2014;196:509–22. doi:10.1534/genetics.113.158220.

53. Kardos M, Luikart G, Bunch R, Dewey S, Edwards W, McWilliam S, et al. Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. Molecular Ecology. 2015;24:5616–32. doi:10.1111/mec.13415.

54. Pavlidis P, Jensen JD, Stephan W, Stamatakis A. A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. Molecular Biology and Evolution. 2012;29:3237–48. doi:10.1093/molbev/mss136.

55. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genetics. 2014;10:e1004148.

56. Snelling WM, Hoff JL, Li JH, Kuehn LA, Keel BN, Lindholm-Perry AK, et al. Assessment of imputation from low-pass sequencing to predict merit of beef steers. Genes. 2020;11:1–16. doi:10.3390/genes11111312.

57. Forni S, Aguilar I, Misztal I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution. 2011;43:1. doi:10.1186/1297-9686-43-1.

58. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Research. 2005;15:1496–502. doi:10.1101/gr.4107905.

59. Cole JB. PyPedal: A computer program for pedigree analysis. Computers and Electronics in Agriculture. 2007;57:107–13.

60. Leroy G, Mary-Huard T, Verrier E, Danvy S, Charvolin E, Danchin-Burge C. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genetics Selection Evolution. 2013;45:1. doi:10.1186/1297-9686-45-1.

61. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. In: Bioinformatics. Oxford University Press; 2018. p. i884–90.

62. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience. 2020;9:1–14. doi:10.1093/gigascience/giaa051.

63. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. http://arxiv.org/abs/1303.3997. Accessed 31 Jul 2019.

64. Picard Toolkit. Broad Institute, GitHub Repository. 2019.

65. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.

66. Pedersen BS, Quinlan AR. Mosdepth: Quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34:867–8. doi:10.1093/bioinformatics/btx699.

67. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

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

69. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19:1655–64.

70. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

71. Yang J, Hong Lee S, Goddard ME, Visscher PM. Genome-Wide Complex Trait Analysis (GCTA): Methods, Data Analyses, and Interpretations. Springer. 2013;:215–36. doi:10.1007/978-1-62703-447-0_9.

72. Weir BS, Cockerham CC. No Title. Evolution. 1984;38. https://pubmed.ncbi.nlm.nih.gov/28563791/. Accessed 22 Oct 2020.

73. Turner S. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of Open Source Software. 2018;3:731. doi:10.1101/005165.

74. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32:1749–51. doi:10.1093/bioinformatics/btw044.

75. Tortereau F, Servin B, Frantz L, Megens HJ, Milan D, Rohrer G, et al. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics. 2012;13:586. doi:10.1186/1471-2164-13-586.

76. Bhati M, Kadri NK, Crysnanto D, Pausch H. Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. BMC Genomics. 2020;21:1–14. doi:10.1186/s12864-020-6446-y.

77. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biology. 2006;4:0446–58.

78. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Research. 2005;15:1566–75.

79. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. American Journal of Human Genetics. 2018;103:338–48.

80. Degiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: Increased sensitivity, robustness and flexibility. Bioinformatics. 2016;32:1895–7.

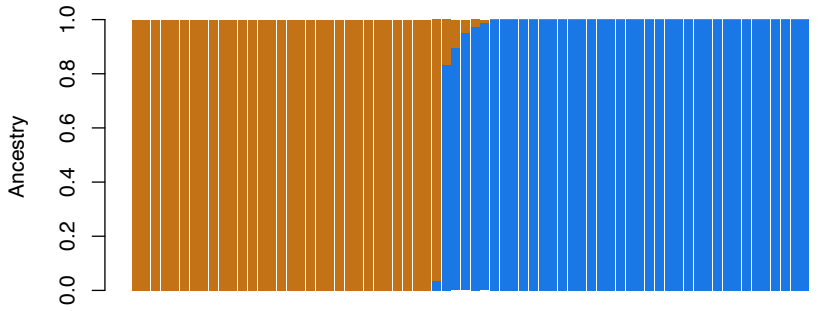
81. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. In: Molecular Ecology Resources. Blackwell Publishing Ltd; 2017. p. 78–90.

82. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30:2503–5. doi:10.1093/bioinformatics/btu314.

2.9 Supplementary files

Supplementary files are available at

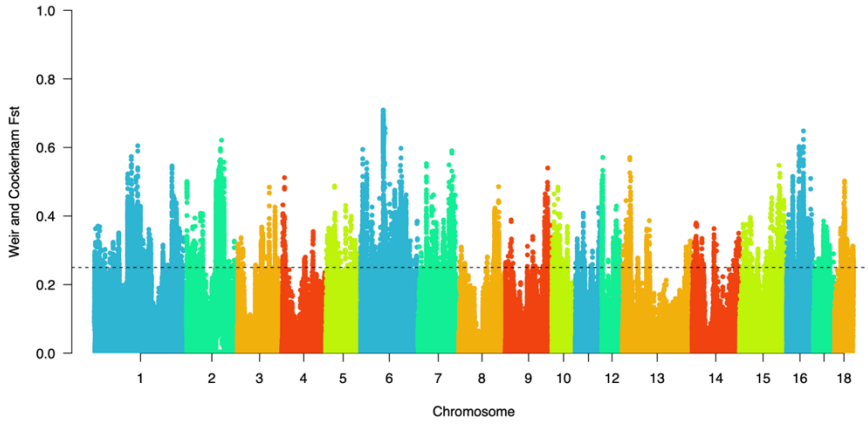
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07610-5#Sec20



Supplementary Fig. S2.1: Admixture analysis.

Ancestry of 70 pigs with K = 2 ancestral populations estimated using 1,207,189 biallelic SNPs after LD pruning. Ancestry proportions were estimated using the ADMIXTURE software. Each bar represents an individual and the colors indicate

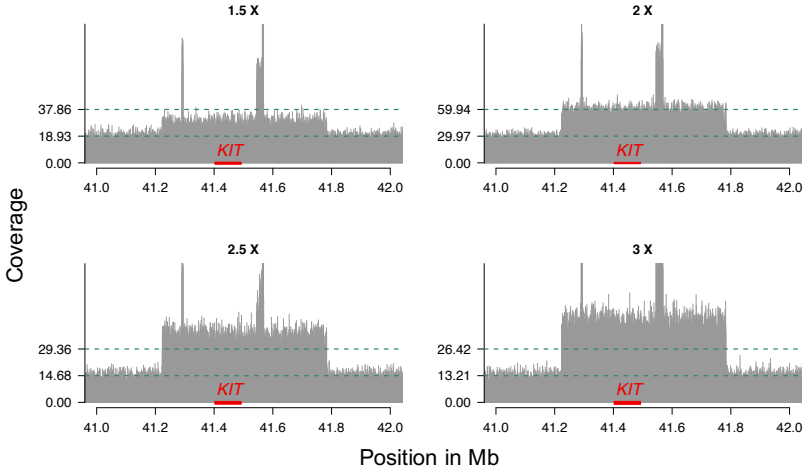
the proportion of genes originating from K ancestral populations. The animals are ordered by population.



Supplementary Fig. S2.2: Manhattan plot of FST values.

Weir and Cockerham FST estimates were calculated in 10 kb sliding windows between 31 dam and 35 sire boars. Black dotted line indicates a value of 0.25.

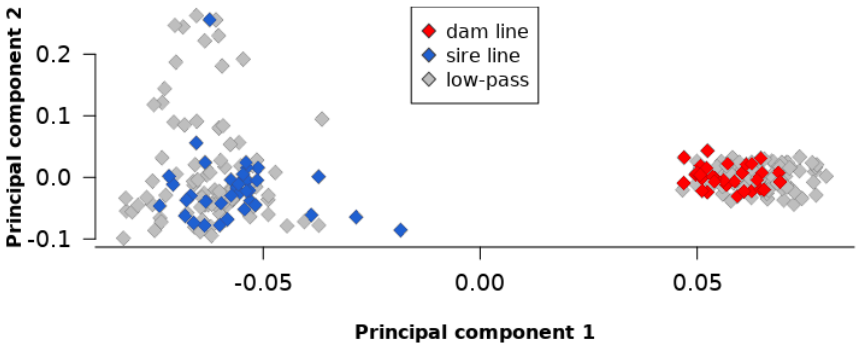
Supplementary Table S2.1: List of variants listed in the OMIA database and their corresponding frequency in the two pig lines.



Supplementary Fig. S2.3: Depth of coverage at a region on SSC8 encompassing the KIT gene.

Representative plots of different depth of coverage detected in the sequenced pigs at a large duplication (DUP1 - SSC8: 41,223,212 - 41,783,660 bp) encompassing the KIT gene. Grey vertical bars represent the absolute coverage observed in four animals. The green dotted lines represent the median and 2*median coverage along SSC8. In order to determine the number of extra copies, we divided for each sequenced animal the average coverage observed at SSC8 by the average coverage observed at DUP1 (chr8: 41,223,212 - 41,783,660 bp). The number of additional copies ranged from 1 to 4. 16, 27, 4, 20, 2 and 1 animal had 1.5 - 1.9x, 2x, 2.1 - 2.4x, 2.5 - 2.9x and 3x the average coverage of SSC8 at DUP1, respectively. The average copy number was 2.07 and 2.19 in the dam and sire line, respectively. The 560 kb duplication (DUP1) encompasses two smaller duplications DUP2 and DUP3/4. DUP2 is 4.3kb long and upstream, while DUP3 and DUP4 are 23kb and 4.3kb duplications downstream of the KIT gene.

Supplementary Table S2. 2: Candidate signatures of selection based on CLR and iHS analyses. Genes annotated to the region are given for each signature of selection.



Supplementary Fig. S2.4: Principal components analysis of key ancestor and low-pass sequenced animals.

Plot of the first two principal components showing the relationship of 96 dam and 96 sire animals sequenced at low (< 1.5-fold) coverage and 32 dam and 38 sire animals sequenced at high (~ 16.5-fold) coverage.

3 Comparison of two multi-trait association testing methods and sequence-based fine mapping of six additive QTL in Swiss Large White pigs

Adéla Nosková[1], Arnav Mehrotra[1], Naveen Kumar Kadri[1], Audald Lloret-Villas[1], Stefan Neuenschwander[2], Andreas Hofer[3], Hubert Pausch[1]

[1] ETH Zürich, Universitätstrasse 2, 8092, Zürich, Switzerland

[2] ETH Zürich, Tannenstrasse 1, 8092, Zürich, Switzerland

[3] SUISAG, Allmend 10, 6204 Sempach, Switzerland

Published in *BMC Genomics, (2023).*

https://doi.org/10.1186/s12864-023-09295-4

Contribution: I participated in conceiving the study, analysing the results, and writing of the manuscript.

3.1 Abstract

3.1.1 Background

Genetic correlations between complex traits suggest that pleiotropic variants contribute to trait variation. Genome-wide association studies (GWAS) aim to uncover the genetic underpinnings of traits. Multivariate association testing and the meta-analysis of summary statistics from single-trait GWAS enable detecting variants associated with multiple phenotypes. In this study, we used array-derived genotypes and phenotypes for 24 reproduction, production, and conformation traits to explore differences between the two methods and used imputed sequence variant genotypes to fine-map six quantitative trait loci (QTL).

3.1.2 Results

We considered genotypes at 44,733 SNPs for 5,753 pigs from the Swiss Large White breed that had deregressed breeding values for 24 traits. Single-trait association analyses revealed eleven QTL that affected 15 traits. Multi-trait association testing and the meta-analysis of the single-trait GWAS revealed between 3 and 6 QTL, respectively, in three groups of traits. The multi-trait methods revealed three loci that were not detected in the single-trait GWAS. Four QTL that were identified in the single-trait GWAS, remained undetected in the multi-trait analyses. To pinpoint candidate causal variants for the QTL, we imputed the array-derived genotypes to the sequence level using a sequenced reference panel consisting of 421 pigs. This approach provided genotypes at 16 million imputed sequence variants with a mean accuracy of imputation of 0.94. The fine-mapping of six QTL with imputed sequence variant genotypes revealed four previously proposed causal mutations among the top variants.

3.1.3 Conclusions

Our findings in a medium-size cohort of pigs suggest that multivariate association testing and the meta-analysis of summary statistics from single-trait GWAS provide very similar results. Although multi-trait association methods provide a useful overview of pleiotropic loci segregating in mapping populations, the investigation of single-trait association studies is still advised, as multi-trait methods may miss QTL that are uncovered in single-trait GWAS.

3.2 Background

Genome-wide association studies (GWAS) combine genotype and phenotype information to identify trait-associated variants. Genotypes at polymorphic loci are tested for association with phenotypes to determine their impact on traits of interest. Multi-trait GWAS can increase the statistical power over single-trait GWAS because they exploit cross-phenotype associations at pleiotropic loci [1–3].

Several methods have been developed to detect pleiotropic variants. These methods can be divided into two groups based on their underlying statistical framework [4, 5]. First, multivariate methods jointly model all traits of interest. This group of methods requires that all individuals included in the study have phenotypic records for all traits analysed, although there are exceptions (e.g., single step GWAS [6], imputation of phenotypes [7]). These methods exploit the genetic covariance between traits, thereby increasing statistical power over their univariate counterparts [5, 8, 9], unless all traits are highly correlated [9, 10]. Second, the meta-analysis of summary statistics enables to combine results from single trait

GWAS, which means that the analyses can be carried out with different sets of individuals for each trait [1, 11–13].

The power to detect trait-associated variants increases as the marker density increases [14–17]. Low-pass sequencing is a cost-effective approach to provide high marker density [18–21]. The imputation from medium-density arrays to the whole-genome sequence level using a sequenced reference panel is another approach to provide sequence variant genotypes for large cohorts. Large and diverse porcine haplotype reference panels facilitate imputing sequence variant genotypes at high accuracy for animals from various breeds [22]. Medium-sized breed-specific reference panels may enable similar accuracy while reducing computational costs [23]. Imputation from medium-density genotypes to the whole-genome sequence level has been explored when high-density array-derived genotypes were not available [16, 24–26].

Only few genome-wide association studies have been conducted in the Swiss Large White (SLW) population. Becker et al. [27] performed association tests between 26 complex traits and 60K SNPs genotyped in 192 breeding boars. This effort revealed only 4 QTL likely because the sample size was too small. Large-scale association testing had been conducted in other pig breeds (e.g., [27, 28, 29]). Fat deposition and weight gain-related traits have been considered frequently in these GWAS as they are economically relevant and highly heritable. Previous GWAS led to tens of proposed candidate genes affecting these traits, including *MC4R*, *BMP2*, *IGF2*, and *CCND2* [31–36].

In this paper, we compare single-trait, multivariate and meta-GWAS in 5,753 genotyped pigs from a Swiss breed to investigate the genetic architecture of 24 traits. We inferred sequence variant genotypes from a

sequenced reference panel to identify candidate causal variants for six pleiotropic QTL.

3.3 Results

3.3.1 Single-trait association studies between array genotypes and 24 traits

A detailed description of the 24 traits considered in our study including their grouping into four categories (reproduction, production, conformation, all) is shown in Table 3.1. Marked pairwise correlations exist between the drEBV of the 24 traits (Additional file 3.3.1). The SNP-based heritability estimates of the drEBV (Table 3.1) were between 0.04 and 0.67.

Table 3.1: Traits with their abbreviations, full descriptions, corresponding trait group and descriptive statistics of the drEBV.

Trait group	Acr.	Full name [unit of raw records]	N_1	N_2	Mean ± SD_2	Min, Max_2	$h^2 \pm SE_1$	FDR
Reproduction	PSP	Stillborn piglets [%]	2,886	2,610	0.4 ± 3.57	-9.46,19.63	0.07 ± 0.01	-
	PUP	Underweight (< 1 kg) piglets [%]	2,886	2,554	0.88 ± 3.20	-13.83,18.29	0.09 ± 0.01	2.50
	NBA	Piglets born alive [number]	2,886	2,697	2.46 ± 1.46	-4.04,8.35	0.13 ± 0.01	-
	GL	Gestation length [days]	2,886	2,866	0.55 ± 0.88	-2.77,5.18	0.29 ± 0.02	5.00
Production	MAS	Meat surface in *longissimus dorsi* [cm^2]	5,457	3,829	-3.71 ± 3.35	-15.17,12.13	0.53 ± 0.02	5.00
	IMF	Intramuscular fat content in MAS [%]	5,422	3,335	0.23 ± 0.53	-1.99,2.92	0.52 ± 0.01	1.67
	DRL	Drip loss [%]	5,515	3,595	0.86 ± 1.57	-6.12,7.32	0.43 ± 0.02	-
	LMC	Lean meat content [%]	5,468	5,155	0.19 ± 1.87	-10.34,8.53	0.46 ± 0.02	0.18
	PH24	pH 24 h postmortem in the loin	4,615	2,729	0 ± 0.05	-0.20,0.20	0.06 ± 0.01	-
	ADFI	Average daily feed intake [kg/day]	5,467	5,109	0 ± 0.15	-0.67,0.62	0.37 ± 0.01	0.10
	DWG	Daily weight gain on test [g/day]	5,467	5,008	10.72 ± 68.15	-256.1,347.5	0.24 ± 0.02	0.15
	LDWG	Lifetime daily weight gain [g/day]	5,468	5,399	14.27 ± 31.06	-113.3,168.0	0.23 ± 0.02	0.11
	MT	Loin muscle thickness [mm]	5,467	3,944	-2.25 ± 1.74	-8.07,5.58	0.09 ± 0.01	5.00
	BFT	Back fat thickness [mm]	5,468	5,120	-0.87 ± 1.95	-8.96,7.7	0.47 ± 0.01	0.25
Conformation	WSFH	Week to steep fetlock, hind legs †	5,434	2,274	0.06 ± 0.19	-0.82,0.81	0.05 ± 0.01	-
	GAIT	Gait	5,434	2,025	0.39 ± 0.22	-0.83,1.45	0.08 ± 0.01	-
	BFL	Bent to pre-bent curve of forelegs †	5,434	2,581	0.14 ± 0.14	-0.55,0.87	0.41 ± 0.01	2.50
	NEIH	Narrowed (reduced) to enlarged inner hoof, hind claws †	5,430	2,564	0.18 ± 0.13	-0.45,0.64	0.06 ± 0.01	-
	SCH	Sword- to chair-legged, hind legs †	5,434	2,361	0.17 ± 0.17	-0.65,0.91	0.06 ± 0.01	-
	XOH	X- to O-legged (knocked-kneed to bow-legged), hind †	5,434	2,516	0.04 ± 0.12	-0.6,0.67	0.04 ± 0.01	-
	CL	Carcass length [cm]	5,433	3,638	4.11 ± 2.66	-5.25,14.64	0.67 ± 0.01	0.17
	NT	Number of teats (both sides)	5,433	5,419	2.19 ± 0.78	-0.72,5.03	0.41 ± 0.01	0.25
	NIT	Number of inverted teats	5,427	2,721	-0.28 ± 0.21	-1,0.96	0.09 ± 0.01	5.00
	NUT	Number of underdeveloped teats	5,433	2,234	0.01 ± 0.09	-0.27,0.45	0.09 ± 0.01	2.50

The false discovery rate (FDR) and genomic heritability based on drEBV (h^2) were based on array genotypes.

1 before filters

2 after filters

† raw phenotypes for conformation traits are recorded on a scale from 1 to 7, where 4 represents the optimum value, but phenotypes for genetic evaluation and hence (deregressed) breeding values reflect the deviation from the optimum

Mixed model-based single-trait genome-wide association studies (stGWAS) between 40,382 SNPs and deregressed estimated breeding values (drEBV) for 24 traits in 5,753 Swiss Large White (SLW) pigs revealed 237 significantly associated variants ($P < 1.24 \times 10^{-6}$). Fifteen out of 24 traits had at least one significantly associated variant (Table 3.1; Additional file 3.3.2; Additional file 3.3.3).

The number of variants that exceeded the Bonferroni-corrected significance threshold was between 1 for GL, NIT, MAS and MTF, and 49 for ADFI (Table 3.1). The inflation factors of the stGWAS were between 0.85 for NT and 1.03 for SCH with an average value of 0.95 ± 0.04 across all 24 stGWAS indicating that population stratification was properly considered.

The 237 associations were detected at 99 unique SNPs located at 11 QTL on SSC1, 5, 11, 15, 17 and 18. The two strongest associations were detected between the number of teats (NT) and a variant on SSC7 (MARC0038565 at 97,652,632 bp, P: 3.35×10^{-35}), and between lifetime daily weight gain (LDWG) and a variant on SSC1 (ASGA0008077 at 270,968,825 bp, P: 3.28×10^{-28}). 57 SNPs were significantly associated with more than one trait (two SNPs, ALGA0123414 and ASGA0008077, were associated with six traits) suggesting that pleiotropic effects are present and detectable in our dataset.

3.3.2 Comparison of multi-trait studies using array genotypes

In order to exploit genetic correlations among the traits to detect pleiotropic loci, we conducted multivariate linear mixed model-based (mtGWAS) association testing and performed a multi-trait meta-analyses

of the single-trait GWAS (metaGWAS[1]) for traits within the four trait categories. For unbiased investigation of differences between both methods, we considered between 1,074 and 2,689 individuals with complete phenotypic records for all traits within a trait category (Table 3.2) for the multi-trait analyses.

Table 3.2: Number of QTL in trait groups revealed by each of the methods.

Group	Reproduction	Production	Conformation	All
Number of traits	4	10	10	24
Number of animals with records for all traits	2,553	1,927	2,689	1,074
Mean correlation between drEBV (± SD)	0.14 ± 0.18	0.32 ± 0.24	0.17 ± 0.19	0.10 ± 0.16
QTL - stGWAS1* (in N traits)	2 (2)	5 (8)	4 (5)	11 (15)
QTL - mtGWAS1*	0	5	3	4
QTL - metaGWAS1*	0	7	3	4
QTL - metaGWAS2*	0	6	3	7
QTL - metaGWAS2 (imputed WGS)	0	6	3	6

[1] performed in pigs with complete records in the trait groups
[2] performed in GWAS conducted in between 2,025 and 5,419 pigs
[*] based on array-derived genotypes at 40,382 SNPs

Both methods yielded similar results, but the metaGWAS[1] revealed more significantly associated variants as well as QTL (Table 3.2; Figure 3.1; Additional files 3.4-3.6). Across the four trait categories, the metaGWAS[1] revealed slightly more significant SNPs (Figure 3.1A) resulting in a 18% smaller FDR than mtGWAS. The metaGWAS[1] revealed 65 unique variants that were significantly associated with at least one trait category, of which 41 were also detected using mtGWAS. The mtGWAS revealed only associations that were also detected by metaGWAS[1] (Figure

3.1B). The P-values of lead SNPs were highly correlated (r = 0.8), but slightly lower (i.e., more significant) in the metaGWAS[1] than the mtGWAS (Figure 3.1C).

Neither of the multi-trait methods detected significantly associated SNP for the reproduction trait category. For the conformation trait category, mtGWAS and metaGWAS[1] revealed 15 and 18 associated SNPs, respectively. The associated SNPs defined three QTL on SSC7, 10, and 17. For the production trait category, the mtGWAS and metaGWAS[1] revealed 26 and 46 associations, respectively. Both methods revealed QTL at SSC1, 16, two QTL at SSC17, and SSC18. The metaGWAS[1] revealed two additional QTL at SSC1 and SSC11. When all 24 traits were combined for 1,074 pigs, the mtGWAS and metaGWAS[1] revealed only 5 and 7 associated SNPs, respectively. These SNPs spanned four QTL on SSC5, 7, 17, and 18.

Seven QTL detected by mtGWAS and metaGWAS[1], were also detected by stGWAS (Figure 3.1D). Both multi-trait methods detected three QTL on SSC11, 16 and 17 that were not detected in the stGWAS. Four QTL detected in the stGWAS were not revealed by either of the multi-trait methods: these were QTL on SSC7, 11, and 15 that were slightly above the Bonferroni-corrected significance threshold for one trait (GL with $P = 5.03 \times 10^{-7}$, NIT with $P = 1.06 \times 10^{-6}$, PUP with $P = 2.66 \times 10^{-7}$, respectively), and one QTL on SSC10 that was associated with MT ($P = 5.60 \times 10^{-7}$) and MES ($P = 8.44 \times 10^{-7}$). From the seven QTL detected by both multi-trait and single-trait methods, six were associated with more than one trait in stGWAS. For the six pleiotropic QTL, at least one single trait analyses revealed more associated variants, and smaller P-value of the top SNP, than the mtGWAS or metaGWAS[1] (Additional file 3.7).

Using summary statistics from stGWAS for a multi-trait metaGWAS facilitates including data from animals with partially missing phenotypes. In order to maximize the power to identify trait-associated pleiotropic variants, we reran the metaGWAS using summary statistics from stGWAS with all available animals per trait (Additional file 3.8), denoted as metaGWAS^2.

Compared to the previous metaGWAS^1 that was based on fewer individuals that had phenotypes for all traits, the number of SNPs exceeding the Bonferroni-corrected significance threshold ($P < 1.24 \times 10^{-6}$) increased by 10, 14 and 48 to 28, 60 and 55 in the conformation, production, and all groups, respectively (Additional file 3.9). No significant markers were detected for the reproduction trait category. Across the four trait groups, the metaGWAS^2 revealed 86 variants, from which 34 were also detected by both other methods, while 21 were detected only by the metaGWAS^1 (Figure 3.1B). Including all available samples into the metaGWAS^2 did not reveal any additional QTL (Figure 3.2A).

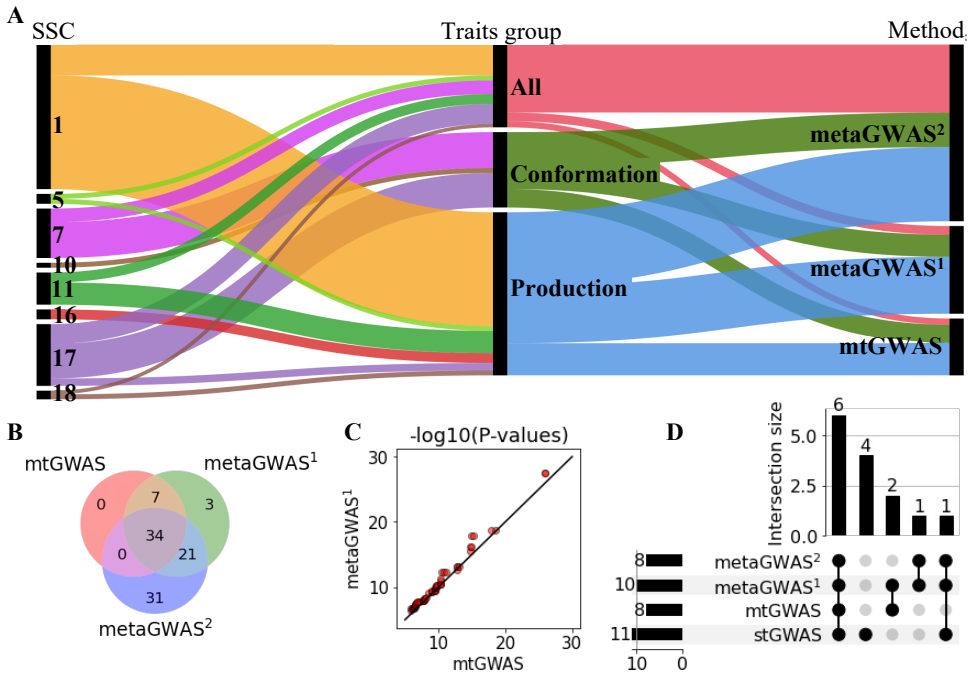


Figure 3.1: Comparison of variants associated with 24 traits from 3 multi-trait GWAS methods. Multivariate (mtGWAS), meta-analyses with complete dataset (metaGWAS[1]), and meta-analyses including samples with missing trait records (metaGWAS[2]) were based on array-derived genotypes. **(A)** Proportions of significantly associated variants discovered across chromosomes, groups of traits and multi-trait methods. **(B)** Overlaps between the associated variants revealed by each of the methods. Sum across all four trait-groups. **(C)** QQ plot between -log10(P) of variants (N = 41) associated in both mtGWAS and metaGWAS[2]. The line denotes a correlation of 1. **(D)** QTL detected by different methods across all trait groups.

3.3.3 Reference panel and imputation of array genotypes to sequence level

Whole-genome sequence-variant genotypes from 421 pigs were used to impute the medium-density genotypes of 5,753 pigs to the whole-genome sequence level. The principal components analysis (PCA; Additional file 3.10) of a GRM built from 16 million biallelic SNP genotypes confirmed that the sequenced reference panel is representative for our GWAS cohort. Five-fold cross-validation indicated high accuracy

of imputation (Additional file 3.11) with values of 0.92, 0.97 and 0.95 for the squared Pearson's correlation (R^2) between true and imputed allele dosages, cross-validated proportion of correctly imputed genotypes (concordance ratio – CR) and model-based accuracies from Beagle5.2 (Beagle DR2), respectively. Although the model-based estimate from Beagle was highly correlated with the Pearson R^2 (0.81), the Beagle DR2 values were consistently higher.

3.3.4 Imputed sequence-based association studies

The stGWAS between 24 traits and the 16,051,635 imputed variants revealed 45,288 variants exceeding the Bonferroni-corrected significance threshold in 7 QTL regions which largely agreed with the results from the array-based GWAS (Additional file 3.12). Two QTL on SSC5 and SSC12 were detected in stGWAS for BFT and NT, respectively, while the other five QTL were associated with multiple traits. The sequence-based GWAS revealed one additional QTL (on SSC12) but did not detect significant association at 5 previously detected QTL likely due to a more stringent Bonferroni-corrected significance threshold resulting from a 350-fold denser marker panel.

A metaGWAS using the summary statistics from stGWAS between the 16,051,635 imputed whole-genome sequence variants and 24 traits using all animals revealed six QTL on SSC1, 5, 7, 17 and 18 (Figure 3.2, Table 3.3) with a total of 9,774 variants exceeding the Bonferroni-corrected significance threshold. When the six lead imputed SNPs were fitted as fixed effects in stGWAS, the peaks in the metaGWAS Manhattan plot disappeared (Additional file 3.13), indicating that the lead SNP accounted for the QTL variance. Four QTL revealed by metaGWAS were significantly associated in stGWAS exclusively with production traits,

whereas two QTL (on SSC7 and SSC17) were associated with traits from both the production and conformation categories (Figure 3.2C). The total trait variance explained per QTL ranged from 0.18 to 11.02 % (Figure 3.2C).

Table 3.3: Imputed lead variants in pleiotropic QTL revealed by a multi-trait meta-analyses.

bp)	Stop (bp)	Lead SNP	P-value	MAF	Ref|Alt	#traits	Candidate gene(s)
172	162,736,434	1_159637589	5.26 × 10-13	0.31	G* | C	3	MC4R
332	272,315,496	1_270599319	2.85 × 10-36	0.14	C | CT*	6	na
537	66,210,538	5_65997650	5.40 × 10-10	0.41	G | A*	1	CCND2
538	99,261,691	7_97636980	8.09 × 10-45	0.34	A* | C	4	VRTN, ABCD4
927	20,928,904	17_15643342	1.19 × 10-84	0.20	C* | T	5	BMP2
756	10,904,242	18_10678235 18_10678293	2.00 × 10-19	0.32	C | A* A | G*	2	na

* Minor allele in the genotyped dataset; Reference alleles were determined according to the Sscrofa11.1 genome assembly.

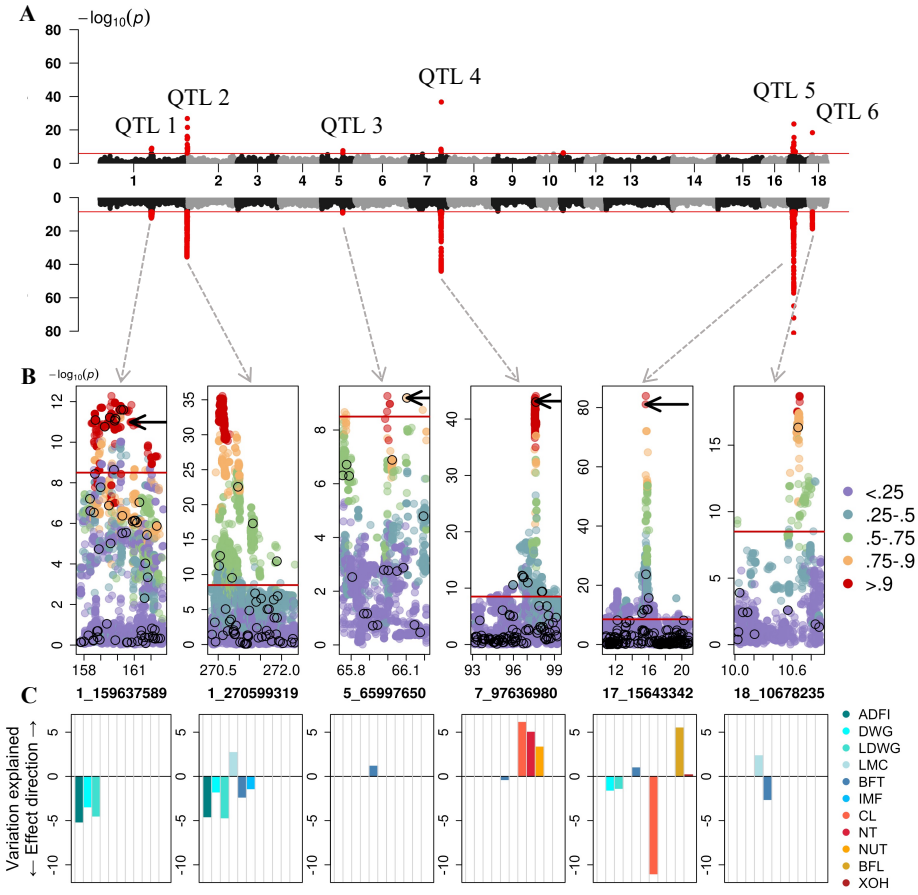


Figure 3.2. Fine mapping of six QTL detected by metaGWAS.

(A) Manhattan plots from array (upper) and imputed sequence (bottom) variants in metaGWAS with 24 traits. **(B)** Linkage disequilibrium between the lead SNPs and all other variants. Black circles mark array SNPs, arrows point to previously proposed causal variants. The red line indicates the genome-wide Bonferroni-corrected significance threshold. **(C)** Variation explained (in % of the drEBV variance) by alternative alleles of the lead SNPs in the single traits. Production traits are in blue scale (ADFI - Average daily feed intake; DWG - Daily weight gain on test; LDWG - Lifetime daily weight gain; LMC - Lean meat content; BFT - Back fat thickness; IMF - Intramuscular fat content in loin), and conformation traits in red scale (CL - Carcass length; NT - Number of teats - both sides; NUT - Number of underdeveloped teats; BFL - Bent to pre-bent, front legs; XOH - X- to O-legged).

3.3.5 QTL 1 with lead SNP 1_159637589

A QTL on SSC1 was between 157.73 and 162.74 Mb and encompassed 34,312 imputed sequence variants including 1,157 that were significantly associated in the sequence-based metaGWAS ($-\log_{10}(P) > 8.5$). The QTL was associated with ADFI, DWG, and LDWG, and explained 5.16, 3.45, and 4.49% of the trait variance, respectively. The lead SNP at this QTL was an imputed intergenic sequence variant (rs692827816) located at 159,637,589 bp ($P = 5.26 \times 10^{-13}$). rs692827816 was in high LD ($R^2 > 0.90$) with 935 variants that had similar P values. One of the variants in high LD was a missense variant (rs81219178) at 160,773,437 bp within the *MC4R* gene, which has previously been proposed as candidate causative variant for growth and fatness traits [37]. rs81219178 segregated at MAF of 0.36 in the SLW population and was imputed from the reference panel with high accuracy (DR2 = 1).

3.3.6 QTL 2 with lead SNP 1_270599319

Another QTL on SSC1 was between 270.33 and 272.32 Mb and encompassed 21,683 imputed sequence variants including 3,765 that were significant in the sequence-based metaGWAS (Additional file 3.14). The QTL was associated with ADFI, DWG, LDWG, LMC, BFT, and IMF, and explained 4.57, 1.77, 4.69, 2.70, 2.35, and 1.40% of the trait variance, respectively. The lead variant was an imputed insertion polymorphism (C>CT) located at 270,599,319 bp ($P = 2.85 \times 10^{-36}$), approximately 13 kb downstream from *ASS1* and 53 kb upstream from *FUBP3*. The lead variant was in high LD ($R^2 > 0.90$) with 747 variants. The MAF of the lead variant was 0.14 and its genotypes were imputed from the reference panel with high accuracy (DR2 = 0.98).

3.3.7 QTL 3 with lead SNP 5_65997650

A QTL on SSC5 was located between 65.75 and 66.21 Mb and encompassed 5,065 imputed sequence variants, from which eight were significantly associated with BFT (Additional file 3.15). The QTL explained 1.16% of the phenotypic variance of BFT. The lead SNP was an imputed sequence variant located in an intergenic region at 65,997,650 ($P = 5.40 \times 10^{-10}$; rs346219461) which was in high LD ($R^2 > 0.90$) with seven other variants. rs346219461 was 7.60 kb downstream the fibroblast growth factor 6 (*FGF6*) encoding gene and it had MAF of 0.41, and it was imputed from the reference panel with high accuracy (DR2 = 0.99). The rs346219461 was in LD ($R^2 = 0.82$) with a non-coding variant (rs80985094 at 66,103,958 bp, $P = 6.41 \times 10^{-10}$) in the third intron of *CCND2*, that was previously proposed as putative causal variant for backfat thickness [36].

3.3.8 QTL 4 with lead SNP 7_97636980

A pleiotropic QTL on SSC7 was between 93.12 and 99.26 Mb and it encompassed 31,013 imputed sequence variants including 2,341 that were significant in the sequence-based metaGWAS. The QTL was associated with BFT from the production group (0.37% variance explained), and CL, NT, and NUT from the conformation group, where it explained 6.11, 5.00, and 3.32% of the trait variance, respectively. The lead SNP was an imputed variant (rs333375257 at 97,636,980 bp, $P = 8.09 \times 10^{-45}$) located 12.7 kb downstream *VRTN*. The rs333375257 had MAF of 0.34 and was imputed from the reference panel with high accuracy (DR2 = 0.99). The rs333375257 was in high LD ($R^2 > 0.90$) with 424 sequence variants. A previously described candidate causal variant (rs709317845 at 97,614,602 bp, $P = 6.71 \times 10^{-44}$) for the number of thoracic vertebrae [38] was in LD ($R^2 > 0.99$) with the rs333375257. In addition, 334 significant variants in

the ATP binding cassette subfamily D member 4 (*ABCD4)* gene were detected. This gene was proposed to impact NT in a Duroc population [39] with top SNP rs692640845 at position 97,568,284. In our study, the rs692640845 was highly significantly associated ($P = 1.49 \times 10^{-42}$) with CL, NT and NUT, and in almost complete LD ($R^2 = 0.98$) with the lead SNP.

3.3.9 QTL 5 with lead SNP 17_15643342

A pleiotropic QTL on SSC17 encompassed 82,132 imputed sequence variants including 2,112 that were significantly associated residing between 10.80 and 20.93 Mb. The QTL was associated with DWG, LDWG, and BFT from the production group (explaining 1.56, 1.36, and 0.97% of the trait variance, respectively), and with BL, BFL, and XOH from the conformation group (explaining 11.02, 5.48, 0.18% of the trait variance, respectively). The strongest association was from an imputed sequence variant (rs342044514 at 15,643,342, $P = 1.19 \times 10^{-84}$) in an intergenic region 106 kb upstream the *BMP2* gene. The variant had MAF of 0.2 and it was imputed from the reference panel with high accuracy (DR2 = 0.97). The lead SNP was in high LD ($R^2 > 0.90$) with two other variants. One of them was a previously proposed candidate causative variant for carcass length [35] (rs320706814 at 15,626,425, $P = 8 \times 10^{-82}$) in an intergenic region upstream of the *BMP2* gene, 17 kb away from the lead SNP.

3.3.10 QTL 6 with lead SNP 18_10678235

A QTL on *SSC18* between 10.03 and 10.90 Mb encompassed 5,790 imputed sequence variants including 408 that were significant. The QTL was associated with LMC and BFT, explaining 2.34 and 2.62% of the phenotypic variance, respectively. The QTL had two lead SNPs ($P = 2 \times$

10^{-19}) in complete LD, which were imputed sequence variants located at 10,678,235 bp (rs338817164) and 10,678,293 bp (rs334203353). Both variants had MAF of 0.32. The imputation accuracy was 0.80. They were in high LD ($R^2 > 0.90$) with other 11 intergenic variants.

3.4 Discussion

Single- and multi-trait genome-wide association studies involving array-derived and imputed sequence variant genotypes from 5,753 SLW pigs enabled us to investigate the genetic architecture of 24 complex traits from three trait groups. The response variables for the association tests were deregressed breeding values because the genotyped pigs had progeny-derived phenotypes. Progeny-derived phenotypes have been frequently used to perform association studies in animals that lack own performance records for the traits of interest. To avoid false-positive associations arising from the accumulation of family information in the progeny-derived phenotypes [40], we used the deregressed breeding values and weighed them according to equivalent relatives' contributions.

The single-trait association analyses revealed 26 trait × QTL associations at eleven QTL of which seven were associated with at least two traits. Exploiting genetic correlations among the traits in a multi-trait framework revealed association for six out of the seven pleiotropic QTL detected in the stGWAS. Despite considering up to 10 phenotypes in the mtGWAS and up to 24 phenotypes in the metaGWAS, the multivariate methods applied in our study revealed three QTL, that were not detected by the single-trait association studies. The multi-trait methods did not reveal association at four QTL that were revealed by stGWAS. These QTL

had low P-values, and perhaps because of higher penalty for multiple testing their effects were too small to be detected in our medium-sized cohort. The phenomenon that associations detected by stGWAS might disappear in multi-trait analyses has been reported earlier [12, 41].

The mtGWAS and metaGWAS[1] detected largely the same associated SNPs for almost all trait-groups. However, the metaGWAS[1] revealed more associated SNPs, and lower P-values for the lead SNPs. Combining all 24 traits in the mtGWAS revealed five associated SNPs, and the metaGWAS[1] conducted with the same individuals detected association of seven SNPs. Multivariate linear mixed models may suffer from over-parametrisation and loss of power when more than ten traits are considered [42]. The low number of detected QTL might also result from low genetic correlations between the 24 traits, or from a small sample size (N = 1,074 pigs with non-missing records for all 24 traits).

The metaGWAS approach enabled us to establish a larger sample size by considering summary statistics from stGWAS that were conducted with a various number of individuals (i.e., some pigs had missing records for some of the traits). In this setting, the number of associated SNPs detected by the metaGWAS[2] increased to 55 (10-fold higher than before). According to Bolormaa et al. [1], in situations where each stGWAS is performed on partially different set of individuals, the metaGWAS approach still appropriately considers variances and covariances among the t-values. It is worth mentioning that there are also frameworks that enable considering samples with partially missing phenotypes in multi-trait GWAS [43–45], but these avenues were not explored in the current study.

Our comparisons between GWAS approaches considered microarray-derived SNP genotypes. The imputation of array-derived genotypes up to the sequence level provides more statistical power to

identify associated loci because causal variants are in the data and directly tested for association with traits of interest [24, 26]. The pigs in our study were genotyped at 44,733 SNPs. No samples were genotyped with denser (e.g., 600K) arrays precluding the stepwise imputation of genotypes up to the sequence level. The imputation of sequence variant genotypes into sparsely genotyped samples may be inaccurate particularly for rare alleles [16]. However, imputation in our mapping cohort with a haplotype reference panel of 421 sequenced animals was accurate. This is likely because the haplotype reference panel mainly contained animals from the target breed.

The imputed sequence variants were used in metaGWAS to fine-map QTL and prioritise candidate causal variants. The top associated variants in four QTL were variants that had been previously identified as candidate causal variants [35–39], indicating that they might also underpin these QTL in the SLW breed. However, none of the proposed candidate causal variants was the top variant in our association studies possibly indicating sampling bias [46], presence of multiple trait-associated variants in linkage disequilibrium [47, 48], or that the top variants were inaccurately imputed [49]. It is also possible that the previously reported candidate causal variants are not causal. Further in-depth functional investigations are required to determine and validate the molecular mechanisms underpinning the QTL identified in our study.

Our meta-analyses approach using imputed sequence variants revealed six QTL of which five were associated with multiple traits. Several porcine pleiotropic loci are underpinned by heterozygous loss-of-function alleles that may have fatal consequences in the homozygous state [50–52]. Pleiotropic QTL have also been described in pigs for highly

correlated traits [32, 53]. The meta-analyses of 24 traits conducted in our study revealed six QTL, that were significantly associated in one to six stGWAS. These results emphasize the importance of the single-trait analyses for dissecting the pleiotropic effects. A QTL on SSC17 which is associated with traits belonging to distinct trait categories. This QTL is associated with carcass length ($P = 1 \times 10^{-62}$) and daily weight gain ($P = 6 \times 10^{-13}$). These two traits are barely correlated with each other (r = 0.02 - 0.05). Another pleiotropic QTL on SSC1 at ~ 270 Mb was associated with six traits ADFI, DWG, LDWG, LMC, BFT, and IMF, that were moderately to highly correlated (mean r ± SD = 0.37 ± 0.27). This chromosomal region harbours QTL for backfat thickness and feed efficiency-related traits in other pig populations [34, 54]. However, candidate causal variants underpinning this QTL had not been proposed so far. The lead SNP in our study was at position 270,599,319 bp, which is 13 kb downstream from the *ASS1* gene. Expression of *ASS1* has been associated with digestive tract development, cell adhesion, response to lipopolysaccharide, and arginine and proline metabolism in pigs [55, 56]. Considering its putative role in energy metabolism, we propose *ASS1* as a positional and functional candidate gene for a pleiotropic QTL at ~ 270 Mb. Further functional annotations of the trait-associated variants in the non-coding regions might help elucidating the genetic mechanism underpinning this pleiotropic QTL.

3.5 Conclusions

Multi-trait associations analyses provide strength of evidence for the presence or absence of a QTL segregating in populations. Here, we compared the multivariate linear model with meta-analyses of single-trait summary statistics using real data. Both approaches performed similarly in correlated groups of traits with complete datasets. The ability of meta-

analyses to include different sets of individuals and unrestricted number of traits promoted the detection power. Thus, we recommend using the meta-analyses for getting overview of pleiotropic QTL in cohorts with more than 10 traits. For analyses of reduced and correlated groups of traits, the choice of the method seems to provide indifferent results. Yet, we stress the importance of the single-trait analyses for accurate interpretation of the pleiotropic effects and for the assignment of the affected traits.

The reference-guided imputation to whole-genome sequence level assigned genotypes to 22 million variants with high accuracy. Putative causal variants found in literature were among the top variants in the fine-mapped QTL. Our analyses provide overview of the QTL affecting economically important traits in Swiss Large White and might serve as catalogue for future research examining the causal variants for complex traits.

3.6 Methods

3.6.1 Animals and phenotypes

Deregressed estimated breeding values (drEBV) with their corresponding degrees of determination (r^2_{drEBV}) and weights (w_{drEBV}) for 24 traits were provided by the Swiss breeding company SUISAG for 5,753 pigs of the SLW breed. Breeding values were estimated using BLUP multiple trait animal models neglecting genomic information and subsequently deregressed according to Garrick et al. [57]. For all our analyses, we considered drEBV which had $r^2_{drEBV} > 0.3$ and were within five standard deviations from the mean values. We considered only traits for which at least 2,000 genotyped animals had records. The final number

of animals with phenotypes was between 2,025 for gait (GAIT) and 5,419 for number of teats (NT). Up to 37 % of the pigs had missing records for at least one trait. The traits were assembled in four trait groups (Table 3.1): reproduction (4 traits), conformation (10 traits), production (10 traits) and all (24 traits).

3.6.2 Genotypes

Microarray-derived genotypes were available for 17,006 pigs from different breeds. The genotypes were obtained with five SNP panels with medium density. There were 2,970 pigs genotyped with the Illumina PorcineSNP60 Bead Chip comprising either 62,163 (v. 1) or 61,565 (v. 2) SNPs; 13,342 pigs were genotyped using customized 60K Bead Chips comprising either 62,549 (v. 1) or 77,119 (v. 2) SNPs; and 546 pigs had genotypes at 68,528 SNPs obtained with the GeneSeek Genomic Profiler (GGP) Porcine 80K array.

We used PLINK (v. 1.9; [58]) to merge the genotypes from the five SNP panels based on the physical positions of the SNPs according to the Sscrofa11.1 assembly [59] of the porcine genome. Then, we performed a quality control for the combined dataset. We retained unique autosomal SNPs that did not deviate from Hardy-Weinberg proportions ($P < 0.00001$), had SNP- and individual-level genotyping rates above 80%, and minor allele frequency (MAF) greater than 0.5%. Finally, sporadically missing genotypes for the resulting 44,733 variants were imputed for 14,292 animals using Beagle (v. 5.0; [60]).

For the array-based GWAS and for the comparison of the multi-trait GWAS methods, we considered 40,382 SNPs that had MAF greater than 5% in 5,753 animals of the SLW breed.

3.6.3 Genomic heritability

We used the Fisher-scoring algorithm implemented in the GREML module of GCTA (v. 1.92.1; [61]) to estimate variance components while considering the inversed weight of drEBV (w_{drEBV}). The genomic relationship matrix was built for 14,292 individuals with 44,733 SNPs, but only up to 5,753 SLW animals were used to estimate genomic heritability.

3.6.4 Imputation to whole-genome sequence level

Whole-genome sequence (WGS) data were available for 421 SLW pigs that had been sequenced at an average read depth of 7.1x, ranging between 2.35x and 37.5x. This panel also included 32 key ancestors of the genotyped SLW pigs that explained a large fraction of the genetic diversity of the current breeding population [19].

Raw sequence data were trimmed and pruned for low-quality bases and reads with default parameter settings of the fastp software (v. 0.20.0; [62]) and subsequently mapped to the Sscrofa11.1 reference genome using the mem-algorithm of the BWA software (v. 0.7.17; [63]). Duplicated reads were marked with the Picard tools software suite (v. 2.25.2; [64]), followed by sorting the alignments by coordinates with Sambamba tool (v. 0.6.6; [65]). The read depth at each genomic position was calculated with the mosdepth software (v. 0.2.2; [66]), considering reads with mapping quality > 10. Variant calling and filtering followed Genome Analysis Toolkit (GATK - v. 4.1.0; [67]) best practice recommendations. Base quality scores were adjusted using the BaseRecalibrator module while considering 63,881,592 unique positions from the porcine dbSNP (v. 150) as known variants. The discovery, genotyping, and filtering of SNPs and INDELs in

the 421 pigs was done using the HaplotypeCaller, GenomicsDBImport, GenotypeGVCFs and VariantFiltration modules of the GATK.

The filtered WGS dataset containing 421 pigs and 22,018,148 variants (18,839,630 SNPs and 3,178,518 INDELs) with MAF greater than 0.01 was used for the imputation of sequence variant genotypes into the array dataset. The reference panel was pre-phased using SHAPEIT4 (v. 4.2; [68]) using the --sequencing parameter. The target array dataset was pre-phased with SHAPEIT4 using the phased sequence data as the reference. Sequence variant genotypes were imputed with Beagle (v. 5.2; [60]) with an effective population size of 50. The effective population size was estimated using SneP [69].

The accuracy of imputation was assessed empirically by five-fold cross-validation in the 421 animals as follows: 40 animals which were sequenced at high coverage (>10x), were used as target panel. The remaining 381 animals served as the reference panel. The SNP density in the target panel was reduced to 44,733 SNP chip genotypes and subsequently imputed to the sequence level based on 381 reference animals as described above. The imputed and actual genotypes of the target samples were compared to derive concordance ratio (CR; proportion of correctly imputed genotypes) and squared correlation (R^2) between imputed and true genotypes.

Relationship between the animals included in the reference panel and the target animals was assessed with a principal components (PC) analysis. First, a genomic relationship matrix (GRM) was built among 14,629 pigs that had 16,387,582 (partially) imputed biallelic SNPs with MAF > 5% (421 reference animals and 14,208 animals with imputed sequence variant genotypes) using GCTA [61]. Then, the first 10 principal components (PC) of the GRM were obtained with PLINK (v. 1.9).

Post-imputation quality control excluded SNPs with MAF < 5%, model-based accuracy of imputation (Beagle DR2) < 0.6, and deviations from Hardy-Weinberg proportions ($P < 10^{-8}$), resulting in a total of 16,051,635 biallelic variants (13,773,179 SNPs and 2,278,456 INDELs) which were used for association analyses and the fine-mapping of QTL in 5,753 SLW pigs.

3.6.5 Single-trait genome-wide association analysis (stGWAS)

Single marker-based GWAS were conducted between 24 traits (see Table 3.1 for more information about the traits and the number of individuals with records) and either 40,382 array-derived or 16,051,635 imputed sequence variant genotypes using the mixed model-based approach implemented in the GEMMA software (v.0.98.5; [70]).

The linear mixed model fitted to the data was in the following form: $y = W\alpha + x\beta + u + \epsilon$, where y is a vector of phenotypes of *n* animals; W is a vector of ones; α is a vector of corresponding coefficients; x is a vector of marker genotypes, coded as 0, 1 and 2 for genotype A_1A_1, A_1A_2 and A_2A_2; β is the effect of the A_2 allele; $u \sim \mathrm{MVN}_n(0, G\sigma_a^2)$ is a random polygenetic effect with G representing the $n \times n$ -dimensional genomic relationship matrix (GRM); σ_a^2 is the additive genetic variance; $\epsilon \sim \mathrm{MVN}_n(0, I\sigma_e^2)$ is a vector of errors, with I representing an identity matrix; and σ_e^2 is the residual variance. MVN_n denotes the n-dimensional multivariate normal distribution.

The centred GRM was calculated with GEMMA (--nk 1) using either array-based or imputed sequence variant genotypes. The P-value of each SNP was estimated by the score test implemented in GEMMA (-lmm 3).

The stGWAS was run with either all available individuals or considering only preselected individuals that had non-missing phenotypes within a group of traits. The markers were separately filtered for MAF > 5%. Thus, in the latter run, the stGWAS for the reproduction traits included 41,242 variants typed in 2,553 samples; the stGWAS for the production traits included 40,557 variants typed in 2,689 samples; the stGWAS for the conformation traits included 41,168 variants typed in 1,927 samples, and the stGWAS for the 24 traits together included 41,152 variants typed in 1,074 samples with non-missing records.

3.6.6 Multi-trait genome-wide association analyses (mtGWAS)

Multi-trait association tests (mtGWAS) were conducted using a multivariate mixed model-based approach implemented in the GEMMA software (v.0.98.5; [42]). The multivariate linear mixed model was parameterised similar to the stGWAS model ($\mathrm{y} = \mathrm{W}\alpha + \mathrm{x}\beta + \mathrm{u} + \epsilon$), except that $\mathrm{y}, \alpha, \mathrm{u}, \epsilon$ are matrices with d (number of traits) columns, and β is a vector with length d. σ_a^2 and σ_e^2 are $d \times d$ symmetric matrices of genetic and environmental variance components, respectively. Because multivariate association testing as implemented in GEMMA requires phenotype data for all individuals and traits, we considered only 2,553, 1,927, 2,689 and 1,074 individuals, respectively, for the reproduction (4 traits), conformation (10 traits), production (10 traits), and all-trait (24 traits) mtGWAS. The GRM, used during the mtGWAS, was the one from the stGWAS.

3.6.7 Meta-analyses multi-trait genome-wide association (metaGWAS)

A multi-trait meta-analysis (metaGWAS) was conducted with the summary statistics from stGWAS as suggested by Bolormaa et al. [1]. Briefly, the t-values for each marker-trait combination were calculated based on the allele substitution effect and corresponding standard error obtained from the stGWAS. The multi-trait χ^2 statistic was subsequently calculated as $\chi^2_{df=d} = t_j^d V^{-1} t_j$, where t is a $j \times d$ matrix of signed t-values at the j^{th} marker across d traits, and V^{-1} is the inversed $d \times d$ variance-covariance matrix.

P-values for the j markers were calculated with pchisq function with d-1 degrees of freedom, as implemented in R. We carried out the meta-analyses with the 24 traits classified into the same four trait categories as in mtGWAS (reproduction, production, conformation, and all). First, to enable unbiased comparison between the metaGWAS and the mtGWAS results, we considered summary statistics obtained from stGWAS based on individuals with complete records within the respective trait-group. Second, to increase the power of the association tests through maximizing the volume of entering information, hence exploiting the benefit of the metaGWAS approach, we used the stGWAS summary statistics based on all available individuals for the trait (the first run). For clarity we denote the first and second meta-analyses as metaGWAS^1 and metaGWAS^2, respectively. The latter approach was repeated for the fine-mapping of the QTL with imputed sequence variant genotypes.

3.6.8 Comparison of the association methods

We used a 5 % Bonferroni-corrected significance threshold (1.24×10^{-6} and 3.11×10^{-9} for array and imputed sequence variant genotypes, respectively) to consider multiple testing. Genomic inflation factors were calculated to compare the distributions of the expected and observed test statistics.

The statistical power was assessed using false discovery rate (FDR). Following Bolormaa et al. [71], the FDR was calculated as $\frac{P * (1-\frac{A}{T})}{\frac{A}{T} * (1-P)}$, where P is the significance threshold (e.g., 1.24×10^{-6} or 3.11×10^{-9}), A is the number of significant variants and T is the total number of variants tested.

3.6.9 Fine mapping of detected QTL

We defined QTL as a region of 1 Mb non-overlapping windows, containing at least one significantly associated marker. The marker with the smallest P-value within a QTL was defined as lead variant. Linkage disequilibrium (LD) between the lead variant and all other variants was calculated with the PLINK (v. 1.9) --r2 command. Variants within QTL were annotated with Ensembl's Variant Effect Predictor (VEP; [72]) tool using local cache files from the Ensembl (release 104) annotation of the porcine genome. The deleteriousness of missense variants was predicted with the SIFT scoring algorithm [73] implemented in VEP.

The proportion of drEBV variance explained by a QTL was estimated with $\frac{2p\,(1-p)\beta^2}{\sigma^2}$, where p is the frequency of the minor allele of the lead SNP and σ^2 is the drEBV variance; β is the regression coefficient of the lead SNP. To avoid overestimating the variance explained by a lead variant, we followed the approach described in Kadri et al. [74] and estimated the

regression coefficients jointly for all QTL from the stGWAS, i.e., the lead variants of those QTL, that were significantly associated in the stGWAS, were fitted as covariates.

3.7 List of abbreviations

EBV: Estimated breeding value; drEBV: Deregressed estimated breeding value; GWAS: Genome-wide association study; INDEL: Insertion and deletions; LD: Linkage disequilibrium; MAF: Minor allele frequency; metaGWAS: meta-analyses GWAS; mtGWAS: Multi-trait GWAS; PCA: Principal component analysis; QTL: Quantitative trait loci; r^2_{drEBV}: Reliability of the drEBV; SD: standard deviation; SLW: Swiss Large White; SNP: Single nucleotide polymorph; SSC: *Sus scrofa* chromosome; stGWAS: Single-trait GWAS; w_{drEBV}: weight of the drEBV; WGS: Whole-genome sequencing;

3.8 References

1. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, et al. A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. PLoS Genet. 2014;10:e1004198.
2. Bonnemaijer PWM, Leeuwen EM van, Iglesias AI, Gharahkhani P, Vitart V, Khawaja AP, et al. Multi-trait genome-wide association study identifies new loci associated with optic disc parameters. Commun Biol. 2019;2:1–12.
3. Yoshida GM, Yáñez JM. Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. BMC Genomics. 2021;22:1–13.
4. Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 2017;7.

5. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28:2540.

6. Legarra A, Vitezica ZG. Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. Genetics Selection Evolution. 2015;47.

7. Wang H, Pei F, Vanyukov MM, Bahar I, Wu W, Xing EP. Coupled mixed model for joint genetic analysis of complex disorders with two independently collected data sets. BMC Bioinformatics. 2021;22.

8. Galesloot TE, van Steen K, Kiemeney LALM, Janss LL, Vermeulen SH. A Comparison of Multivariate Genome-Wide Association Methods. PLoS One. 2014;9:e95923.

9. Porter HF, O'Reilly PF. Multivariate simulation framework reveals performance of multi-trait GWAS methods. Sci Rep. 2017;7:1–12.

10. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44:1066.

11. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. BIOINFORMATICS APPLICATIONS NOTE. 2010;26:2190–1.

12. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018;50:229–37.

13. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. Am J Hum Genet. 2015;96:21.

14. Goddard ME, Hayes BJ. Genomic Selection Based on Dense Genotypes Inferred From Sparse Genotypes. Proc Assoc Advmt Anim Breed Genet. 2009;18.

15. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.

16. Van Den Berg S, Vandenplas J, Van Eeuwijk FA, Bouwman AC, Lopes MS, Veerkamp RF. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. Genetics Selection Evolution. 2019;51:1–13.

17. Iso-Touru T, Sahana G, Guldbrandtsen B, Lund MS, Vilkki J. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genet. 2016;17.

18. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res. 2021;:266486.120.

19. Nosková A, Bhati M, Kadri NK, Crysnanto D, Neuenschwander S, Hofer A, et al. Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. BMC Genomics. 2021;22:290.

20. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53:120–6.

21. Zhang Z, Ma P, Zhang Z, Wang Z, Wang Q, Pan Y. The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in Duroc pigs. Genomics. 2022;114:340–50.

22. Ding R, Savegnago R, Liu J, Long N, Tan C, Cai G, et al. Nucleotide resolution genetic mapping in pigs by publicly accessible whole genome imputation. bioRxiv. 2022;:2022.05.18.492518.

23. Lloret-Villas A, Pausch H, Leonard AS. Size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle. bioRxiv. 2023;:2023.01.13.523894.

24. Casu S, Usai MG, Sechi T, Salaris SL, Miari S, Mulas G, et al. Association analysis and functional annotation of imputed sequence data within genomic regions influencing resistance to gastro-intestinal parasites detected by an LDLA approach in a nucleus flock of Sarda dairy sheep. Genetics Selection Evolution. 2022;54:1–16.

25. Ros-Freixedes R, Valente BD, Chen C-Y, Herring WO, Gorjanc G, Hickey JM, et al. Rare and population-specific functional variation across pig lines. Genetics Selection Evolution. 2022;54:1–16.

26. Yan G, Liu X, Xiao S, Xin W, Xu W, Li Y, et al. An imputed whole-genome sequence-based GWAS approach pinpoints causal mutations for complex traits in a specific swine population. Science China Life Sciences 2021. 2021;:1–14.

27. Becker D, Wimmers K, Luther H, Hofer A, Leeb T, Moore S. A Genome-Wide Association Study to Detect QTL for Commercially Important Traits in Swiss Large White Boars. PLoS One. 2013;8.

28. Broekema R v, Jonkers IH, Bakker OB. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biol. 2020;10.

29. Johnsson M, Jungnickel MK. Evidence for and localization of proposed causative variants in cattle and pig genomes. Genetics Selection Evolution. 2021;53:1–18.

30. Pig QTL Database. https://www.animalgenome.org/cgi-bin/QTLdb/SS/summary. Accessed 24 Nov 2021.

31. Cai Z, Christensen OF, Lund MS, Ostersen T, Sahana G. Large-scale association study on daily weight gain in pigs reveals overlap of genetic factors for growth in humans. BMC Genomics. 2022;23:1–13.

32. Delpuech E, Aliakbari A, Labrune Y, Fève K, Billon Y, Gilbert H, et al. Identification of genomic regions affecting production traits in pigs divergently selected for feed efficiency. Genetics Selection Evolution. 2021;53:49.

33. Ding R, Zhuang Z, Qiu Y, Ruan D, Wu J, Ye J, et al. Identify known and novel candidate genes associated with backfat thickness in Duroc pigs by large-scale genome-wide association analysis. J Anim Sci. 2022;100:1–8.

34. Gozalo-Marcilla M, Buntjer J, Johnsson M, Batista L, Diez F, Werner CR, et al. Genetic architecture and major genes for backfat thickness in pig lines of diverse genetic backgrounds. Genetics Selection Evolution. 2021;53:1–14.

35. Li J, Peng S, Zhong L, Zhou L, Yan G, Xiao S, et al. Identification and validation of a regulatory mutation upstream of the BMP2 gene associated with carcass length in pigs. Genetics Selection Evolution. 2021;53:1–13.

36. Oliveira HC, Lopes MS, Derks MFL, Madsen O, Harlizius B, van Son M, et al. Fine Mapping of a Major Backfat QTL Reveals a Causal Regulatory Variant Affecting the CCND2 Gene. Front Genet. 2022;0:1241.

37. Kim KS, Larsen N, Short T, Plastow G, Rothschild MF. A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. Mammalian Genome. 2000;11:131–5.

38. Duan Y, Zhang H, Zhang Z, Gao J, Yang J, Wu Z, et al. VRTN is required for the development of thoracic vertebrae in mammals. Int J Biol Sci. 2018;14:667–81.

39. Zhuang Z, Ding R, Peng L, Wu J, Ye Y, Zhou S, et al. Genome-wide association analyses identify known and novel loci for teat number in Duroc pigs using single-locus and multi-locus models. BMC Genomics. 2020;21.

40. Ekine CC, Rowe SJ, Bishop SC, de Koning DJ. Why breeding values estimated using familial data should not be used for genome-wide association studies. G3: Genes, Genomes, Genetics. 2014;4:341–7.

41. Pausch H, Emmerling R, Schwarzenbacher H, Fries R. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genetics Selection Evolution. 2016;48:1–9.

42. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11:407–9.

43. Jiang J, Zhang Q, Ma L, Li J, Wang Z, Liu JF. Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. Heredity 2015 115:1. 2015;115:29–36.

44. Hormozdiari F, Kang EY, Bilow M, Ben-David E, Vulpe C, McLachlan S, et al. Imputing Phenotypes for Genome-wide Association Studies. Am J Hum Genet. 2016;99:89.

45. Aguilar I, Legarra A, Cardoso F, Masuda Y, Lourenco D, Misztal I. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genetics Selection Evolution 2019 51:1. 2019;51:1–8.

46. Barendse W. The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics. 2011;12:1–12.

47. Bickel RD, Kopp A, Nuzhdin S v. Composite Effects of Polymorphisms near Multiple Regulatory Elements Create a Major-Effect QTL. PLoS Genet. 2011;7:e1001275.

48. Abell NS, DeGorter MK, Gloudemans MJ, Greenwald E, Smith KS, He Z, et al. Multiple causal variants underlie genetic associations in humans. Science (1979). 2022;375:1247–54.

49. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genetics Selection Evolution. 2017;49:1–14.

50. Flossmann G, Wurmser C, Pausch H, Tenghe A, Dodenhoff J, Dahinten G, et al. A nonsense mutation of bone morphogenetic protein-15 (BMP15) causes both infertility and increased litter size in pigs. BMC Genomics. 2021;22:1–9.

51. Derks MFL, Lopes MS, Bosse M, Madsen O, Dibbits B, Harlizius B, et al. Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. PLoS Genet. 2018;14:e1007661.

52. Fujii J, Otsu K, Zorzato F, de Leon S, Khanna VK, Weiler JE, et al. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science (1979). 1991;253:448–51.

53. Bovo S, Mazzoni G, Bertolini F, Schiavo G, Galimberti G, Gallo M, et al. Genome-wide association studies for 30 haematological and blood clinical-biochemical traits in Large White pigs reveal genomic regions affecting intermediate phenotypes. Scientific Reports 2019 9:1. 2019;9:1–17.

54. Jiao S, Maltecca C, Gray KA, Cassady JP. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: II. Genomewide association. J Anim Sci. 2014;92:2846–60.

55. Piórkowska K, Żukowski K, Ropka-Molik K, Tyra M, Gurgul A. A comprehensive transcriptome analysis of skeletal muscles in two Polish pig breeds differing in fat and meat quality traits. Genet Mol Biol. 2018;41:125–36.

56. Ponsuksili S, Murani E, Brand B, Schwerin M, Wimmers K. Integrating expression profiling and whole-genome association for dissection of fat traits in a porcine model. J Lipid Res. 2011;52:668.

57. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution. 2009;41:55.

58. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

59. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience. 2020;9:1–14.

60. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103:338–48.

61. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

62. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. In: Bioinformatics. Oxford University Press; 2018. p. i884–90.

63. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997. 2013.

64. Picard Toolkit. Broad Institute, GitHub Repository. 2019.

65. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.

66. Pedersen BS, Quinlan AR. Mosdepth: Quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34:867–8.

67. Depristo MA, Banks E, Poplin R, Garimella K v., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–501.

68. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nature Communications 2019 10:1. 2019;10:1–10.

69. Barbato M, Orozco-terWengel P, Tapio M, Bruford MW. SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. Front Genet. 2015;6 MAR:109.

70. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet. 2012;44:821.

71. Bolormaa S, Hayes BJ, Savin K, Hawken R, Barendse W, Arthur PF, et al. Genome-wide association studies for feedlot and growth traits in cattle. J Anim Sci. 2011;89:1684–97.

72. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17.

73. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–82.

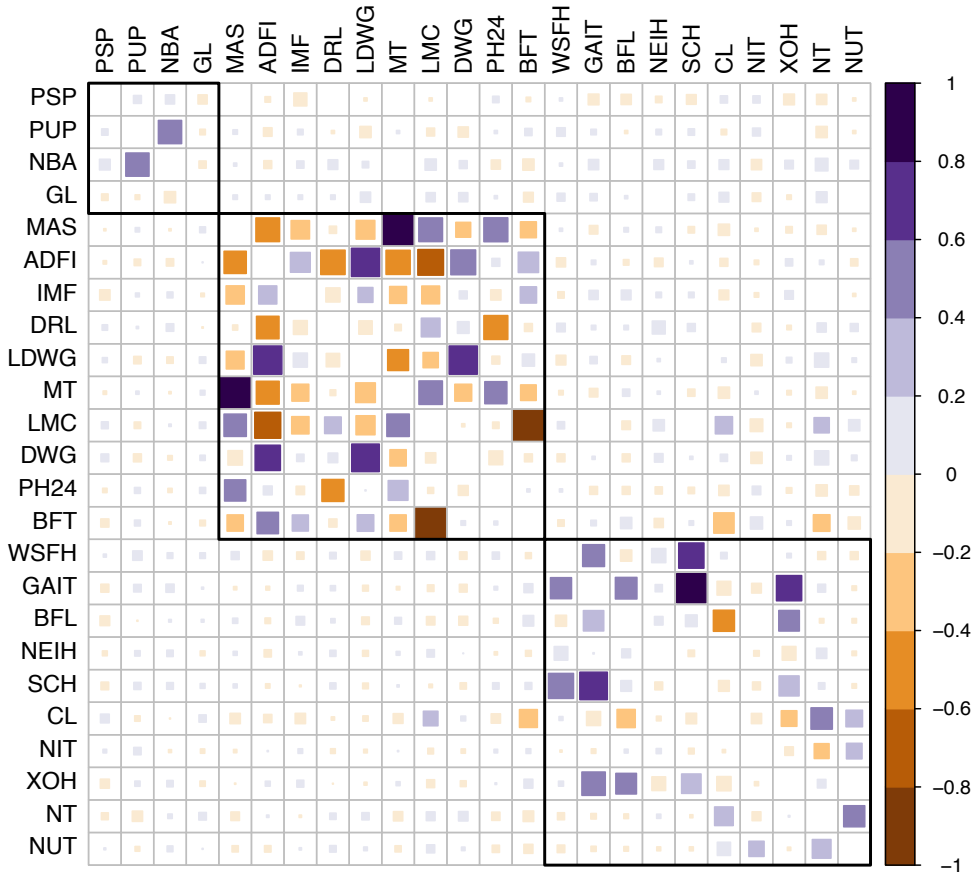
74. Kadri NK, Harland C, Faux P, Cambisano N, Karim L, Coppieters W, et al. Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle. Genome Res. 2016;26:1323–32.

3.9 Additional files

Additional files are available at

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-023-09295-4#Sec25

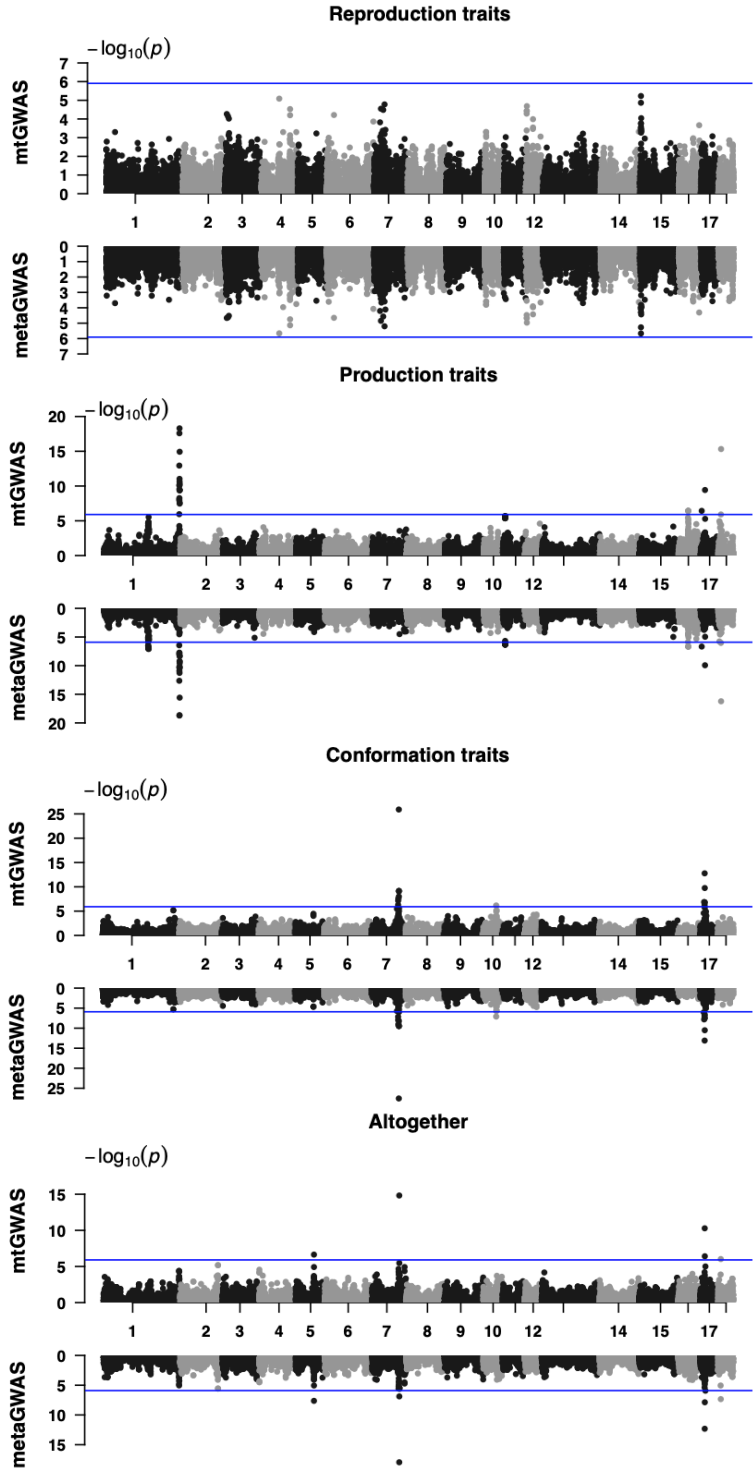
Additional file 3.1: Correlations between the 24 traits.
In the upper and lower triangles are correlations between signed t values, and between deregressed breeding values, respectively. Shades of purple and orange indicate positive and negative correlation coefficients, respectively. The three rectangulars are defining reproduction, production, and conformation groups of traits.



Additional file 3.2: Manhattan plots of single-trait GWAS based on array genotypes.
Each page contains 4, 10 and 10 plots for traits from reproduction, conformation and production groups, respectively. Blue suggestive line is at 5.9.

Additional file 3.3: Results of single-trait GWAS based on array genotypes.
Table contains numbered QTL regions, chromosome, start and stop positions of the QTL, lowest P-value, top SNP with MAF, number of significantly associated variants within the QTL and the single-trait abbreviation.

Additional file 3.4: Manhattan plots of multi-trait GWAS (upper) and meta-analyses GWAS (bottom) based on array genotypes.
Suggestive line is at 5.9.



Additional file 3.5: Results of multi-trait GWAS based on array genotypes.
Table contains numbered QTL regions, chromosome, start and stop positions of the QTL, lowest P-value, top SNP with MAF, number of significantly associated variants within the QTL and the trait-group.

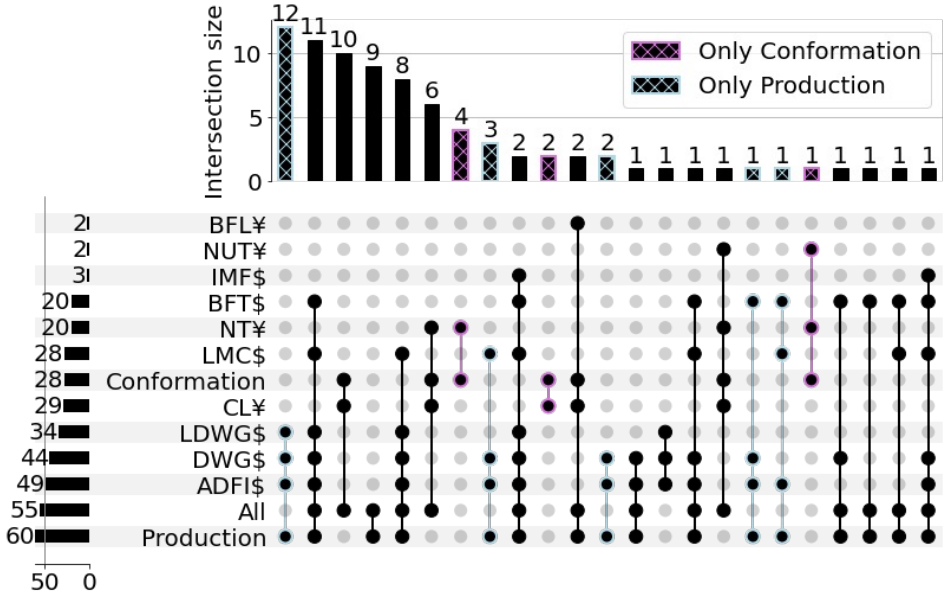
Additional file 3.6: Results of meta-analyses GWAS based on array genotypes and complete dataset.
Table contains numbered QTL regions, chromosome, start and stop positions of the QTL, lowest P-value, top SNP with MAF, number of significantly associated variants within the QTL and the trait-group.

Additional file 3.7: Comparison of pleiotropic QTL across methods based on array genotypes.
Summary table of P-values and number of significantly associated variants within the 6 pleiotropic resulting from single-trait GWAS, meta-analyses GWAS with complete datasets and multi-trait GWAS.

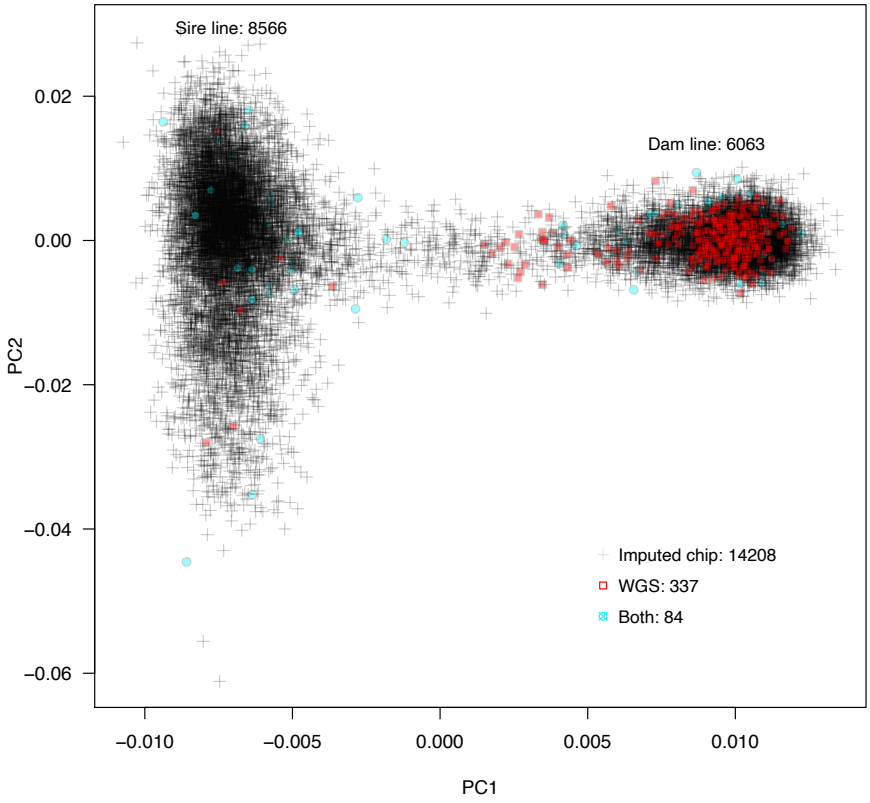
Additional file 3.8: Results of meta-analyses GWAS based on array genotypes and dataset including missing phenotypic records.
Table contains numbered QTL regions, chromosome, start and stop positions of the QTL, lowest P-value, top SNP with MAF, number of significantly associated variants within the QTL and the trait-group.

Additional file 3.9: Number of significantly associated pleiotropic variants in the meta-analyses GWAS within groups using all possible samples and stGWAS of individual traits.

The groups are denoted as $ (production group) and ¥ (conformation group). The single traits are: BFL - Bent to pre-bent curve of forelegs; NUT - Number of underdeveloped teats; IMF - Intramuscular fat content in MAS; BFT - Back fat thickness; NT - Number of teats (both sides); LMC - Lean meat content; CL - Carcass length; LDWG - Lifetime daily weight gain; DWG - Daily weight gain on test; ADFI - Average daily feed intake.

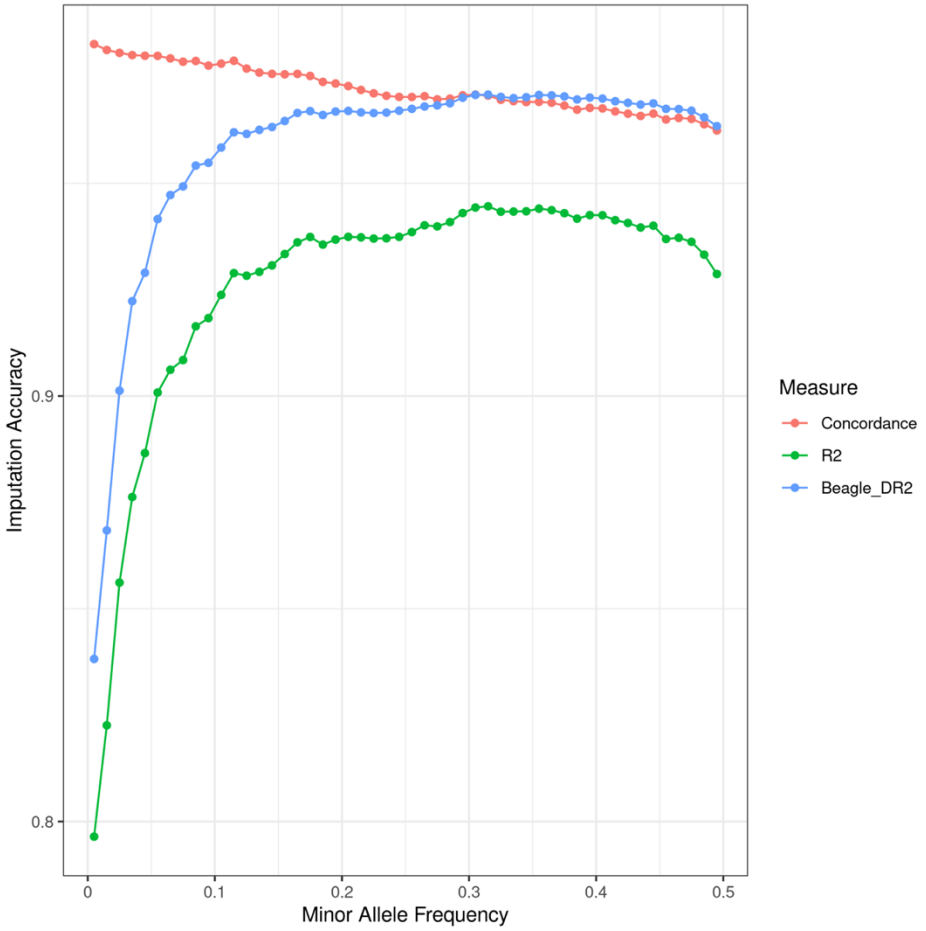


Additional file 3.10: PCA plot of reference panel samples (red and blue) and the target samples with array-genotypes (black +).
First and second principal components captured 7.55 and 0.83% of the genomic



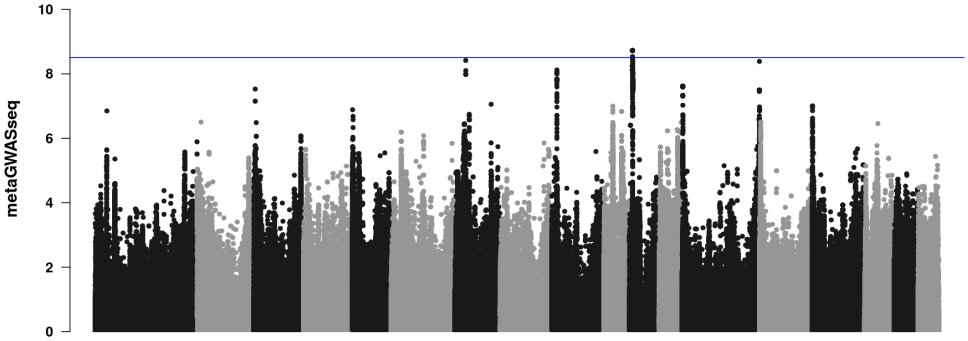
variation.

Additional file 3.11: Accuracy of imputation to whole-genome sequence versus minor allele frequency.
Dosage R2 values from Beagle, and empirical measures of concordance and accuracy (R2) derived in 5 cross-validation sets, were assessed in 421 animals based on 44,733 downsampled chip genotypes.



Additional file 3.12: Results of single-trait GWAS based on imputed sequence variants.
Table contains numbered QTL regions, chromosome, start and stop positions of the QTL, lowest P-value, top SNP with MAF, number of significantly associated variants within the QTL and the single-trait abbreviation.

Additional file 3.13: Manhattan plot of multi-trait meta-analyses GWAS of all 24 traits after fitting the 6 QTL as covariates, which resulted in loss of the peaks.
Suggestive line is at 8.5.



Additional file 3.14: Results of meta-analyses GWAS based on imputed sequence variants and dataset including missing phenotypic records.
Table contains numbered QTL regions, chromosome, start and stop positions of the QTL, lowest P-value, top SNP with MAF, number of significantly associated variants within the QTL and the trait-group.

Additional file 3.15: List of genes ID and genes symbols within the 6 pleiotropic QTL.

4 Exploiting public databases of genomic variation to quantify evolutionary constraint on the branch point sequence in 30 plant and animal species

Adéla Nosková [a], Chao Li [a,b], Xiaolong Wang [b], Alexander S. Leonard [a], Hubert Pausch [a], Naveen Kumar Kadri [a]

[a] Animal Genomics, ETH Zürich, Universitätstrasse 2, 8092, Zürich, Switzerland

[b] International Joint Agriculture Research Center for Animal Bio-Breeding, Ministry of Agriculture and Rural Affairs/Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China

Submitted.

Contribution: I participated in conceiving the study, analysing the results, and writing of the manuscript.

4.1 Abstract

The branch point sequence is a degenerate intronic heptamer required for the assembly of the spliceosome during pre-mRNA splicing. Disruption of this motif may promote alternative splicing and eventually cause phenotype variation. Despite its functional relevance, the branch point sequence is not included in most genome annotations. Here, we predict branch point sequences in 30 plant and animal species and attempt to quantify their evolutionary constraints using public variant databases. We find an implausible variant distribution in the databases from 16 of 30 examined species. Comparative analysis of variants from whole-genome sequencing shows that biased or erroneous variants that are widespread in public databases cause these irregularities. We then investigate evolutionary constraint with largely unbiased public variant databases in 14 species and find that the fourth and sixth position of the branch point sequence are more constrained than coding nucleotides. Our findings show that public variant databases should be scrutinized for possible biases before they qualify to analyze evolutionary constraint.

4.2 Introduction

Precursor messenger RNA (pre-mRNA) splicing is executed by the spliceosome, a large ribonucleoprotein complex that assembles at the intron-exon boundary [1]. Intronic features involved in the recognition and assembly of the spliceosome include the splice sites, polypyrimidine tract and branch point sequence (BPS). A degenerate heptamer containing the branch point residue constitutes the BPS [2]. This motif resides within 50 bases upstream of the 3′ splice site in most of the introns across all

eukaryotes. The heptamer includes highly conserved thymine and adenine residues, but the adenine itself acts as branch point during pre-mRNA splicing [3–5].

Mutations in BPS can promote alternative splicing and manifest phenotype variation [6]. However, despite their functional relevance, BPS are not readily accessible in most gene transfer files. Lack of experimentally proven branch points [7, 8] and the degenerate nature of the sequence encompassing the branch point complicate systematic annotation of this regulatory motif.

Computational methods have been developed to predict BPS [4, 9–11]. Recently, Kadri et al. [12] quantified evolutionary constraint on computationally predicted BPS in the human and bovine genomes using exhaustive variant catalogues established from whole-genome sequencing (WGS). Their analyses showed that the BPS encompasses evolutionarily conserved thymine and adenine residues that are more strongly depleted for variants than coding sequences, suggesting that they are under extreme purifying selection. Recovery of strongly constrained nucleotides within predicted BPS also shows that the motif can be localized *in silico* with high accuracy. Recent analyses in human genomes suggest that the fourth nucleotide of the heptamer might be more strongly depleted for variants than the branch point itself [13, 14]. However, it remains an open question if this constraint pattern is consistent across evolutionarily distant species.

Here we predict BPS in 30 plant and animal species and attempt to study their constraint using public variant databases. We uncover implausible variant distributions in 16 out of 30 public databases precluding such a study in all species. Investigation of variability of the BPS using unbiased public databases of genomic variation reveals strong

evolutionary constraints on both the branch point and on the position two base pairs upstream in 14 species investigated.

4.3 Results

Purifying selection against deleterious mutations manifests as a depletion of variation overlapping constrained nucleotides. Our previous study showed that constraints can be quantified at nucleotide resolution by counting the number of variable sites within functional classes of annotations [12]. We hypothesized that this approach is applicable to validate and quantify evolutionary constraints on the branch point sequence (BPS) for any species for which an annotated reference genome and a large and unbiased variant database are available.

4.3.1 Bovine public variant database is biased

We conducted a proof-of-concept study with 89,118,442 biallelic SNPs from the bovine EVA database (release 4; [15]) to investigate evolutionary constraints on BPS in the bovine genome. We calculated nucleotide-wise constraint - hereafter referred to as 'variability' - for each position of the BPS relative to the average genome-wide density of variants per 100 bp. Contrary to findings in a catalogue of variants established through WGS [12], the branch point was the least constrained nucleotide in the heptamer when variants from the public database were used (Figure 1C). Implausible constraint patterns were also evident for other well annotated features of the genome. For instance, we found intergenic regions to be less variable than coding regions (Figure 4.1A), and excessive variability at the splice sites (Figure 4.1B). These findings suggested that the bovine EVA database contains biased or erroneous variants.

In fact, the proportion of coding variants was 4-fold higher in the bovine EVA database than a variant catalogue established from WGS (Table 4.1). An excess of coding variants in the EVA database indicated it contains variants discovered from exome sequencing. However, an implausibly low transition to transversion (Ti/Tv) ratio of 0.55 for variants overlapping exons also showed that the database is contaminated with false-positive variants. These irregularities were mainly due to a large batch of variants (n = 38,008,641) from one submitter that included many (7.45%) coding variants with very low Ti/Tv ratio (0.45). Most of these variants (83%; Supplementary Figure S4.1) were unvalidated, i.e., they were not confirmed by another submission. Once all variants private to this batch were removed, a subset of 57,875,698 SNPs largely recovered the expected variability of the investigated features (Figure 4.1DEF). However, high variability of nucleotides overlapping the 5' splice site and a relatively low Ti/Tv ratio (1.69) suggested that this subset is still biased. We repeated the analyses with 34,551,781 variants that were submitted to EVA at least twice. Variants within this subset had a Ti/Tv ratio of 2.20. These variants recovered a pattern of evolutionary constraint that matched previous findings from WGS-derived variants (Figure 4.1GHI;Table 4.1). However, this subset contained fewer coding variants (0.49%) than the WGS-derived catalogue which suggests that strict filtration removed true rare coding variants from the data.

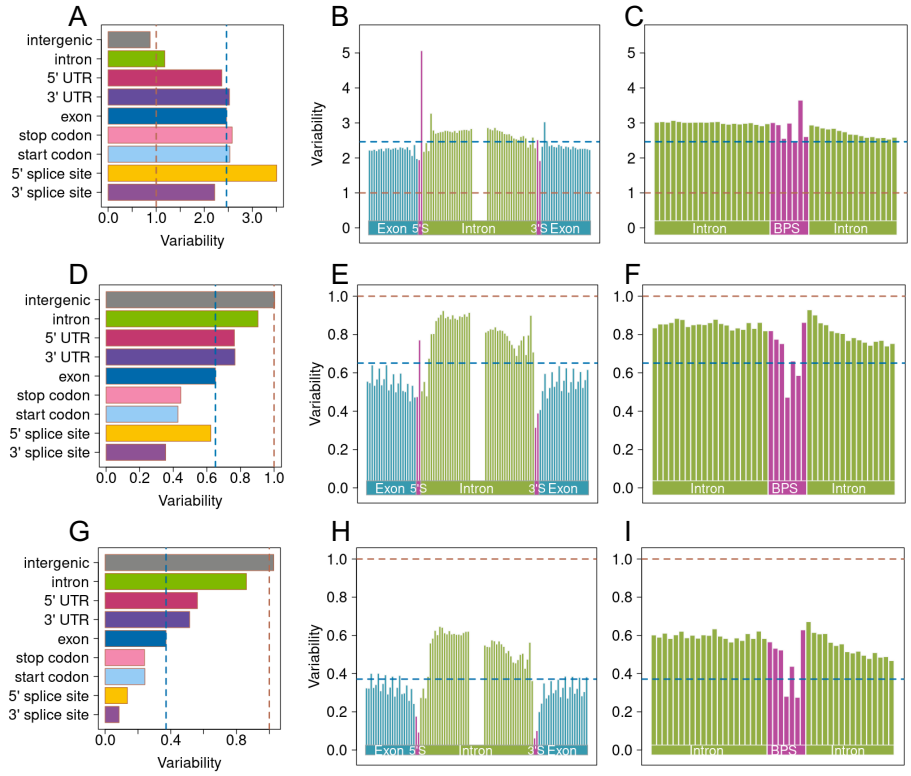


Figure 4.1: Variation in genomic features quantified through a public bovine variant database.

(A, D, G) Variability of nine bovine genomic features. Nucleotide-wise constraint in and around the splice-sites **(B, E, H)** and predicted branch point sequences **(C, F, I).** Constraint was quantified relative to average genome-wide variability using **(A-C)** all 89,118,442 SNPs, **(D-F)** a subset of 57,875,698 SNPs that did not contain variants only submitted by the COFACTOR_GENOMICS_CFG20140112 project, and **(G-I)** a subset of 34,551,781 SNPs that contained only SNPs that were submitted at least twice. Red and blue lines denote average genome-wide and exome variability, respectively.

Table 4.1: Datasets used for analysis.

Species	N	SNPs (Ti/Tv)	% Coding (Ti/Tv)	Predicted BPS	Average variability	Ref. genome	Source
Cow	266	29,227,950 (2.21)	0.91 (3.00)	179,476	1.18	ARS-UCD1.2	[12]
Cow	-	89,118,442 (1.09)	3.27 (0.55)	179,476	3.58	ARS-UCD1.2	EVA rs4
Cow (filtered)	-	34,551,781 (2.20)	0.49 (3.33)	179,476	1.39	ARS-UCD1.2	EVA rs4
Cow (all COF.)	-	34,580,719 (0.55)	7.35 (0.45)	179,476	1.39	ARS-UCD1.2	EVA rs4
Pig	139	24,074,441 (2.36)	0.73 (3.28)	192,744	1.04	Sscrofa11.1	this study
Pig	-	56,883,886 (1.96)	0.91(2.38)	192,744	2.47	Sscrofa11.1	Ensembl variation
Sheep	161	16,198,123 (2.59)	0.54 (4.35)	168,025	0.60	Oar_ramb_v1.0	[16]
Sheep	-	48,541,784 (2.45)	0.82(3.53)	168,025	1.80	Oar_ramb_v1.0	Ensembl variation
Goat	157	13,364,058 (2.48)	0.66 (3.95)	174,407	0.53	ARS1	[17]
Goat	-	31,517,363 (2.40)	0.68(3.88)	174,407	1.26	ARS1	Ensembl variation

4.3.2 Public variant databases reveal expected constraints in pig, sheep, and goat

We quantified constraint patterns in the same genomic features of three additional species to investigate if other public variant databases suffer from similar biases. These analyses were performed in pig, sheep, and goat for which exhaustive variant catalogues were available from both WGS and public databases. The predicted BPS in 192,744 pig, 168,025 sheep, and 174,407 goat introns had a “nnyTrAy” consensus sequence that contained two conserved nucleotides at the branch point itself (position 6) and two bp upstream (position 4) (Figure S4.2A). The branch points were at a median distance of 26 bp upstream of the 3’ splice site (Figure S4.2B).

Most BPS had a canonical ‘TnA’ motif at position 4-6 (88, 90 and 89% in pigs, sheep, and goats, respectively).

Variant discovery in WGS data of 139 pigs, 161 sheep and 157 goats yielded 24,074,441, 16,198,123 and 13,364,058 biallelic SNPs with Ti/Tv ratios between 2.36 and 2.59 (Table 4.1). Variability in the genomic features differed as expected and confirmed previously established patterns of constraint (Figure 2). We observed a striking depletion of variation on the positions 4 and 6 of the predicted BPS. The constraint on both nucleotides was stronger than on coding sequences but did not differ significantly between them (Fisher’s exact test P-values 0.06, 0.79, 0.50 for pig, sheep and goat, respectively).

The public pig, sheep, and goat databases contained more than twice the number of variants we established through WGS but between 28% and 36% overlapped between the databases and WGS for the respective species (Table 4.1). Because the public databases aggregate variant information from many individuals from multiple breeds, the Ti/Tv ratios, proportion of coding variants, and constraints in functional features differed from those established with the smaller WGS subset but were within plausible ranges (Table 4.1; Figure 4.2). Nucleotide-wise constraint in the BPS was also consistent with the pattern obtained from variants established through WGS (Figure 4.2E). As observed with variants from WGS, the constraint did not differ between positions 4 and 6 (Fisher’s exact test P-values 0.48, 0.67, 0.47 for pig, sheep and goat, respectively).

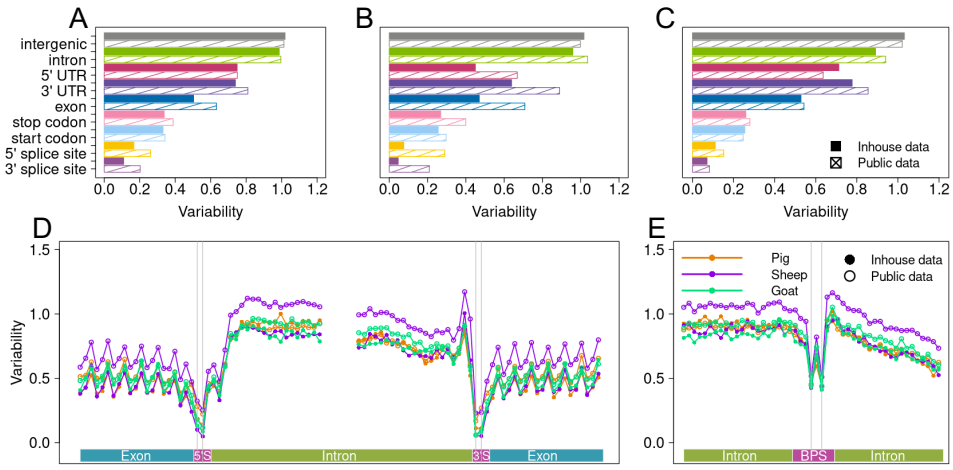


Figure 4.2: Variation in pig, sheep, and goat genomic features quantified through variants from whole-genome sequencing and public databases.

Variability of the nine features of the **(A)** pig, **(B)** sheep, and **(C)** goat genomes. Nucleotide-wise variation relative to average genome-wide variability in and around **(D)** splice sites and **(E)** branch point sequence.

4.3.3 Variant bias is widespread in public databases

Exhaustive variant information from four public databases confirmed evolutionary constraints that were similar to those established from WGS. This encouraged us to conduct a comparative analysis of constraint on the BPS in 26 additional species (18 animals from 13 orders and 8 plants from 4 orders) for which at least one million SNPs were available through EVA (n = 8) or Ensembl (n = 18) databases. We evaluate the quality of these databases prior to the comparative constraint analyses to ensure they are free from erroneous and biased variants.

Two public databases were excluded prior to the comparative analysis because variant density was too low. Variants from 13 databases were incongruent with properties of genome-wide variants and thus were not suitable for an unbiased comparative assessment of evolutionary constraint across species (Table S1, File S1). The variability in intergenic

regions was lower than the average genome-wide variability in 12 excluded databases possibly indicating biased variant distribution due to exome sequencing. An excess of exonic variants in five of these databases is further evidence that the variants fail to represent genome-wide variability. The well-established constraint at the four positions overlapping the 3' and 5' splice sites was absent in five databases (File S1).

Other variant characteristics such as the proportion of coding variants or the Ti/Tv ratio were not abnormal for many excluded databases, indicating that these parameters are not suitable to assess the plausibility of such databases. For instance, a Ti/Tv ratio of 1.96 and 1.11% coding variants in the *Equus caballus* database are compatible with expectations for genome-wide variants (Table S1). Yet, variants from this database revealed an implausibly high variability at both splice sites and downstream the BPS at intronic positions overlapping the polypyrimidine tract (File S1). A plausible Ti/Tv ratio (1.87) and percentage of coding variants (2.89%) may pretend that the *Gallus gallus* variant database is representative for whole-genome variability in chicken. However, an excess of variability of the nucleotides overlapping the 5' splice sites is implausible. Variants from this database also uncovered a constraint pattern in the BPS which deviates from what we established in unbiased variant catalogues.

Only 11 public variant databases (File S1, Table S1) fulfilled our criteria, i.e., they contained on average at least one variant per 1000 base pairs, variant density was higher in intergenic regions than genome-wide (Figure 4.3B) and constraints on the 3' and 5' splice sites were evident (Figure 4.3A). These databases contained between 4,486,640 and 69,472,724 variants of which between 0.61% and 16.51% overlapped

coding sequences. We subsequently conducted a comparative analysis of BPS variation in these species.

A vast majority of the predicted BPS for these 11 species contained the canonical 'TnA' motif overlapping positions 4-6 of the heptamer (between 90% in *Pan troglodytes* and 98% in *Phaseolus vulgaris*; Table S1, Figure S4.3). The predicted branch points were primarily between 14 and 145 bp upstream of the 3' splice site with median distance of 27 bp (Figure S4.4), consistent with BPS placement in other species. The comparative analysis of BPS variation in these 11 species revealed strong constraint on positions 4 and 6 (Figure 3C, Table S1). The constraint between these two positions differed significantly in four of the 11 species investigated. In three of those four species, the constraint was stronger on the position 4 than on the position 6 (Bonferroni corrected Fisher's exact $P < 4.5 \times 10^{-3}$).

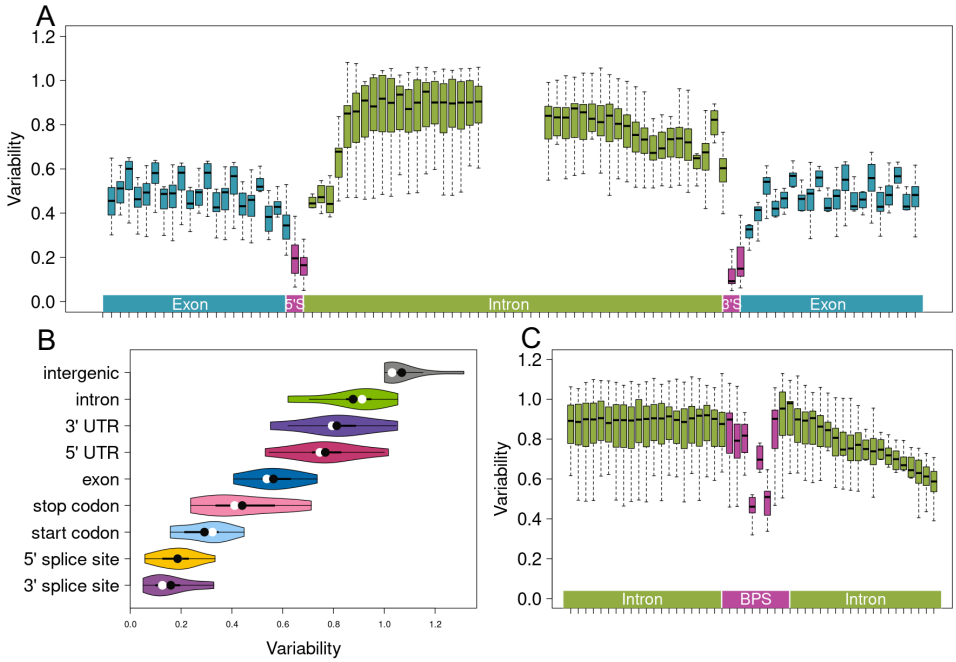


Figure 4.3: Variation in genomic features across 11 species quantified using public variant databases.

Boxplots of nucleotide-wise variability relative to average genome-wide variability in and around **(A)** splice sites and **(C)** predicted branch point sequences. **(B)** Violin plots of variability in nine genomic features. Means and medians are indicated with black and white circles, respectively.

4.4 Discussion

Our comparative analysis of branch point sequence (BPS) variation relied on computationally predicted BPS because exhaustive catalogues of experimentally proven branch points were not available for the species considered. While the length of the reported consensus sequence encompassing the branch point varies from five bases in humans [18] to ten bases in plants [5], all BPS in our study were heptamers that contained the branch point at the 6th position. Constraints on positions 4 and 6 were

striking in all species investigated, corroborating a pivotal role of both residues for spliceosome assembly during pre-mRNA splicing [7]. However, we did not find conclusive evidence for higher constraint on position 4 of the heptamer, as reported for human BPS [13, 14]. Our findings confirm that an adenine residue is the most preferred branch point, and that thymine is the most preferred residue at 2 bp upstream of the branch point [19]. High evolutionary conservation of these nucleotides across distant eukaryotic species reiterates the need to consider them in search for trait-associated variants.

We use public variant databases to assess evolutionary constraints on different genomic features. However, contamination of these databases with erroneous or biased variants can cause flawed interpretation of results [20–22]. Our study corroborates that variants from public databases need to be evaluated carefully due to their partly unknown origin and lack of curation [20, 22–25]. Strict filtration, such as the removal of variants that were submitted only once, was required to recover expected constraint patterns from a public bovine variant database. This approach is only possible with accompanying metadata, which is not always available. Moreover, this approach also removes true rare variants enriched in evolutionary constraint signatures, and as such is not generally advisable.

Assessment of constraint patterns in annotated genomic features is more useful to evaluate the quality of variant databases than inspecting other widely used parameters such as Ti/Tv ratio. We show that the constraint on the splice sites and the proportion of variants in intergenic regions are the most informative for such an assessment. By using a simple and straightforward approach of counting variable sites overlapping genomic features, we show that erroneous and biased variants contaminate 16 of the 30 investigated public variant databases. Since these constraint

patterns are so widespread, such an approach may provide quality assessment for existing or even purely predictive annotations.

4.5 Material and methods

4.5.1 Whole-genome sequence variant databases

We used whole-genome sequencing (WGS) data of 139 pigs that were sequenced at an average read depth greater than 8x. Sequence reads were processed and aligned against the Sscrofa11.1 reference sequence as detailed in [26]. We called variants with DeepVariant (version 1.3.0, [27, 28]), producing a gVCF file per sample. The gVCF files were then merged and filtered using GLnexus (version 1.4.1, [29]) with the DeepVariantWGS configuration, followed by imputation with Beagle 4.1 [30]. Sequence variants were called previously for 266 cattle, 161 sheep and 157 goats (Table 4.1). We considered only biallelic sequence variants for our analyses.

4.5.2 Public variant databases

We downloaded reference sequences and their annotations including non-coding RNAs, as well as a VCF file with polymorphic positions for 30 species from EVA (release 4, [15]), Ensembl (release 107, [31]), or Ensembl Plants (release 55, [32]). Access information for all data is provided in Table S2.

We used these data to evaluate the number of variants, proportion of variants in protein-coding regions, average genome-wide variability (in variants per 100 bases) and transition to transversion (Ti/Tv) ratio. Variants overlapping exons, start codons and stop codons were considered as coding variants.

4.5.3 Prediction of branch point sequences

We followed the approach of Kadri et al. [12] to predict BPS in 30 species using BPP [9]. In short, we obtained coordinates of introns in protein-coding genes from GTF files of each species, mindful of gene-strand orientation. We used species-specific weighted octanucleotide frequencies estimated as suggested by Zhang et al. [9] and the position weight matrix of predicted human BPS for model training [9]. For the analysis of constraint, we only considered the most probable BPS within each intron.

The variability between positions 4 and 6 of the heptamer was compared for each species with Fisher's exact test. We applied Bonferroni-correction to account for multiple testing (number of species tested).

4.5.4 Variation in genic features

Variability was calculated as the number of variants per 100 bp divided by the respective species' genome-wide variability for variants overlapping nine annotated genomic features (3' and 5' splice sites, start and stop codons, 3' and 5' UTR, introns, exons, intergenic regions) and predicted BPS. Genome-wide variability, i.e., average number of variants per 100 bp, was calculated as total number of variants divided by the size of the genome. The genome size was the total length of all chromosomes considered but undetermined bases ("N") were excluded.

4.5.5 Genomic variant database analysis

Based on the analyses of WGS datasets we established three criteria to assess the quality of public databases. The criteria were (i) genome-wide variability of minimum 1 variable site per 1000 bp; (ii) variability in intergenic regions above the average genome-wide; (iii) depletion of variation at the 4 bases overlapping splice sites. Databases not fulfilling all

criteria were excluded from further analyses (Table S1, File S1). Eleven species that satisfied these criteria were considered to estimate constraint patterns.

4.6 References

1. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem. 2015;84:291.
2. Keller EB, Noon WA. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc Natl Acad Sci U S A. 1984;81:7417.
3. Taggart AJ, Lin CL, Shrestha B, Heintzelman C, Kim S, Fairbrother WG. Large-scale analysis of branchpoint usage across species and cell lines. Genome Res. 2017;27:639–49.
4. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 2008;18:88.
5. Zhang X, Zhang Y, Wang T, Li Z, Cheng J, Ge H, et al. A Comprehensive Map of Intron Branchpoints and Lariat RNAs in Plants. Plant Cell. 2019;31:956.
6. Královičová J, Lei H, Vořechovský I. Phenotypic consequences of branch point substitutions. Hum Mutat. 2006;27:803–13.
7. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. Genome Res. 2015;25:290.
8. Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. 2018;32:577–91.
9. Zhang Q, Fan X, Wang Y, Sun M-A, Shao J, Guo D. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics. 2017;33:3166–72.
10. Paggi JM, Bejerano G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. RNA. 2018;24:1647–53.
11. Signal B, Gloss BS, Dinger ME, Mercer TR. Machine learning annotation of human branchpoints. Bioinformatics. 2018;34:920–7.

12. Kadri NK, Mapel XM, Pausch H. The intronic branch point sequence is under strong evolutionary constraint in the bovine and human genome. Communications Biology 2021 4:1. 2021;4:1–13.

13. Zhang P, Philippot Q, Ren W, Lei W Te, Li J, Stenson PD, et al. Genome-wide detection of human variants that disrupt intronic branchpoints. Proc Natl Acad Sci U S A. 2022;119:e2211194119.

14. Blakes AJM, Wai HA, Davies I, Moledina HE, Ruiz A, Thomas T, et al. A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. Genome Med. 2022;14.

15. Cezard T, Cunningham F, Hunt SE, Koylass B, Kumar N, Saunders G, et al. The European Variation Archive: A FAIR resource of genomic variation for all species. Nucleic Acids Res. 2022;50:D1216–20.

16. Li C, Chen B, Langda S, Zhou S, Kalds P, Zhang K, et al. Comparative Genomic Analyses Shed Light on the Genetic Control of High-Altitude Adaptation in Sheep. Submitted. 2023.

17. Li C, Wu Y, Chen B, Cai Y, Guo J, Leonard AS, et al. Markhor-derived Introgression of a Genomic Region Encompassing PAPSS2 Confers High-altitude Adaptability in Tibetan Goats. Mol Biol Evol. 2022;39.

18. Gao K, Masuda A, Matsuura T, Ohno K. Human branch point consensus sequence is yUnAy. Nucleic Acids Res. 2008;36:2257.

19. Zhuang Y, Goldstein AM, Weiner AM. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proc Natl Acad Sci U S A. 1989;86:2752–6.

20. Khafizov K, Ivanov M V., Glazova O V., Kovalenko SP. Computational approaches to study the effects of small genomic variations. Journal of Molecular Modeling 2015 21:10. 2015;21:1–14.

21. LaDuca H, Farwell KD, Vuong H, Lu HM, Mu W, Shahmirzadi L, et al. Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. PLoS One. 2017;12.

22. Maffucci P, Bigio B, Rapaport F, Cobat A, Borghesi A, Lopez M, et al. Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. Proc Natl Acad Sci U S A. 2019;116:950–9.

23. Johnston JJ, Biesecker LG. Databases of genomic variation and phenotypes: existing resources and future needs. Hum Mol Genet. 2013;22:R27–31.

24. Musumeci L, Arthur JW, Cheung FSG, Hoque A, Lippman S, Reichardt JKV. Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies. Hum Mutat. 2010;31:67.

25. Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics. 2004;20:1022–32.

26. Nosková A, Bhati M, Kadri NK, Crysnanto D, Neuenschwander S, Hofer A, et al. Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. BMC Genomics. 2021;22:290.

27. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology 2018 36:10. 2018;36:983–7.

28. Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics. 2021;36:5582–9.

29. Lin MF, Dnanexus OR, Penn J, Bai X, Reid JG, Krasheninina O, et al. GLnexus: joint variant calling for large cohort sequencing. https://doi.org/10.1101/343970.

30. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.

31. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.

32. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. Methods Mol Biol. 2016;1374:115–40.

4.7 Supplementary files

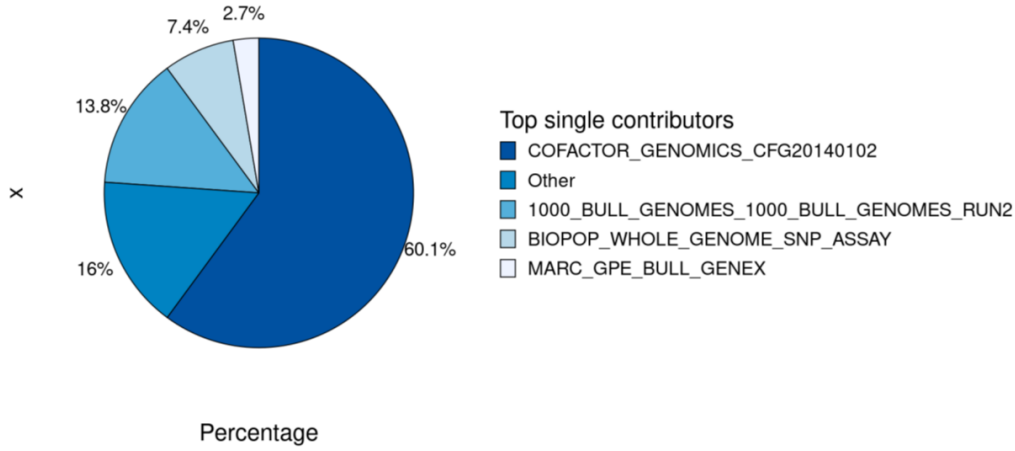


Figure S4.1: Main contributors of single-entry variants.

Top four contributors submitted 84 % of all (N = 51,359,349) singletons. The largest batch (N = 31,580,941) of unvalidated variants was submitted by COFACTOR_GENOMICS_CFG20140112.

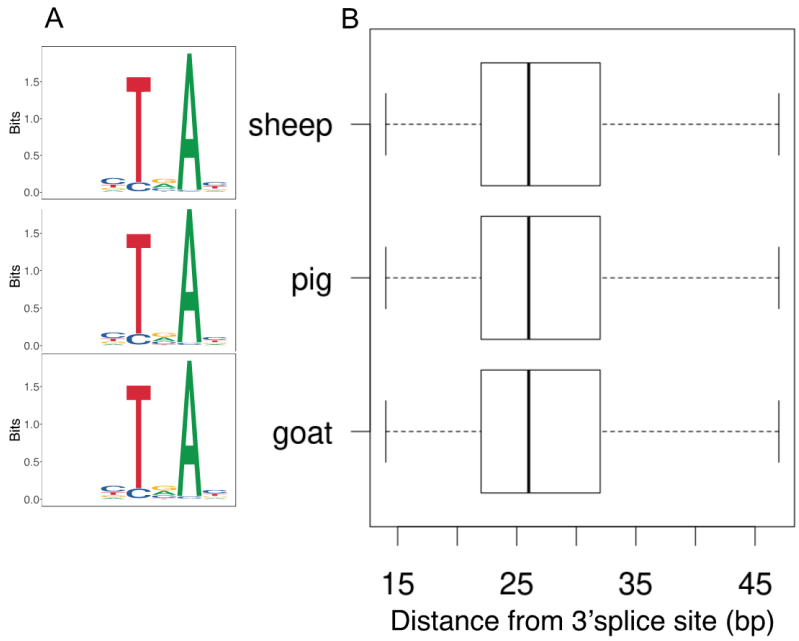


Figure S4.2: Predicted branch point sequences in sheep, pig and goat.

(A) Motif logos of predicted branch point consensus sequences and (B) their placement (distance from 3' splice site in base pairs) in the sheep, pig and goat genomes.

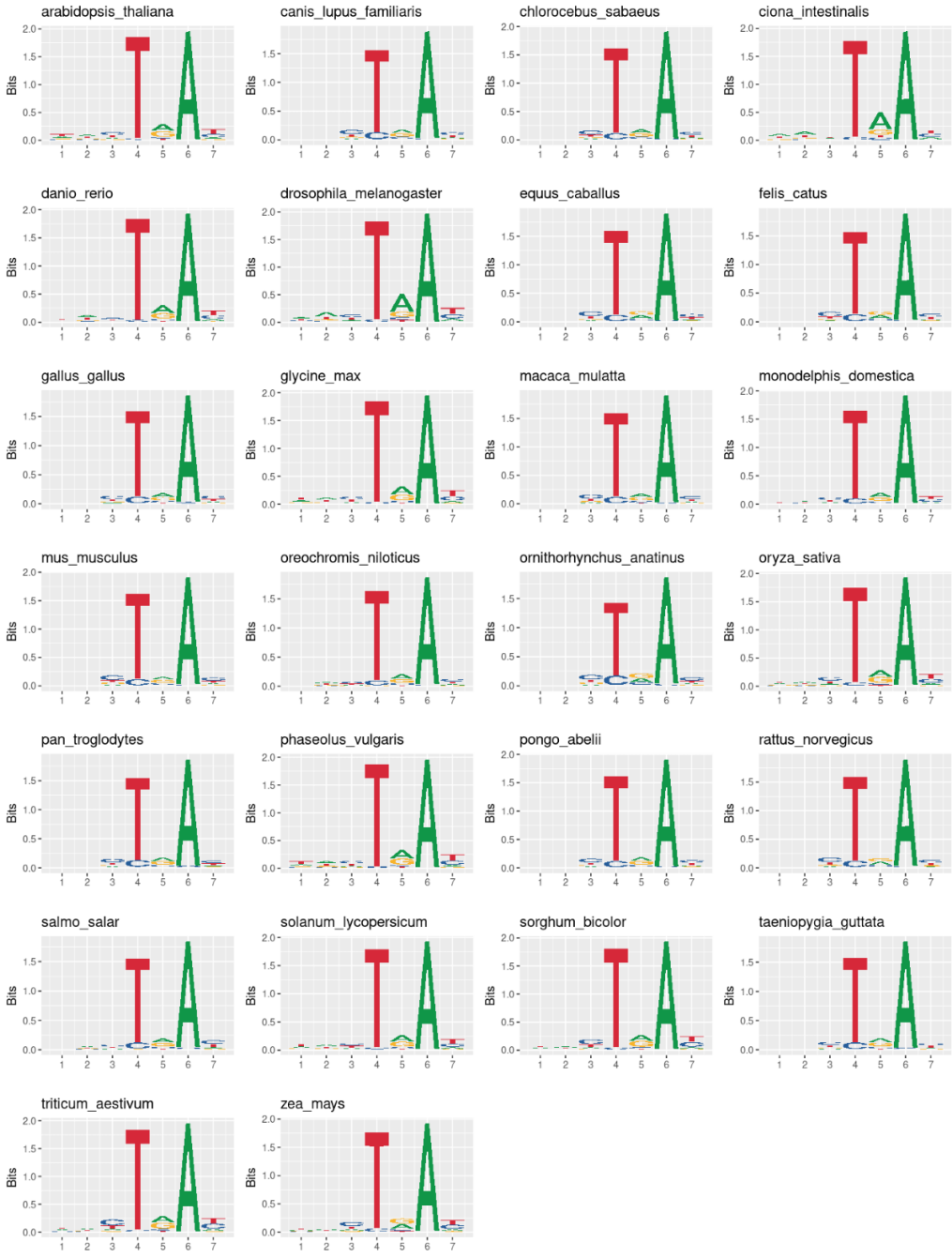


Figure S4.3: Predicted branch point consensus sequences in 26 species.

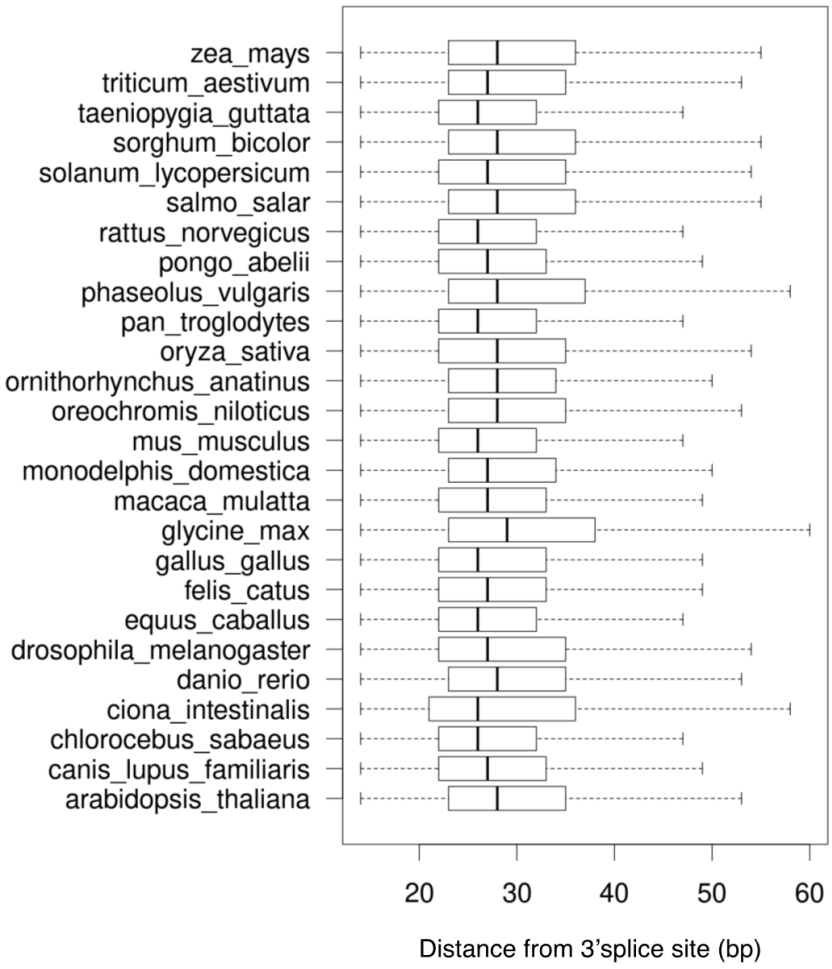


Figure S4.4: Placement of predicted branch points in 26 species.

5 General discussion

This thesis investigated two populations of the Swiss Large White (SLW) pig breed at the whole-genome sequence (WGS) level. Analyses of between- and within-population genetic diversity showed that the two lines are genetically distant enough to be considered as separated populations, despite each originating from the same ancestral breed. Genome-wide association studies (GWAS) of the dam line revealed quantitative trait loci (QTL) affecting economically important traits, such as fattening traits (e.g., average daily feed intake or back fat thickness) and conformation traits (e.g., carcass length or number of teats). The sequence-derived variants from 35 dam line key ancestors enabled imputation from array-derived genotypes up to the WGS level with sufficient accuracy to pinpoint candidate causal mutations. These results widely rely on proper structural and functional annotation of the porcine reference genome. However, the current annotation is missing important genic features, such as branch point sequences, for which this thesis now also provides the predicted positions (Supplementary Table S5.1).

Furthermore, this thesis explored several methods with potential use in breeding programs beyond the pig industry. These approaches were mainly focused on increasing the number of variants through imputation. GLIMPSE-based imputation was shown to enable implementation of low-pass sequencing into populations with existing whole-genome haplotype panels. Its favorable cost facilitates the sequencing of large cohorts of animals when compared to the higher costs for high-throughput sequencing, and thus opens the door for shifting from routine array- to sequencing-based genotyping. Direct imputation from array genotypes to

the WGS level was also explored, which allowed the fine-mapping of multi-trait QTL. To consider the scenario where no variant set would be available from whole-genome sequencing, public variant databases were investigated for various species while intending to enhance current annotation files with prediction of branch point sequences. This approach turned out to be incorrect for many species and highlights the need for critical examination of public variant databases prior to use in genomic studies.

The results of this thesis can be directly utilized by the Swiss pig breeding industry and can serve as a base for a follow-up research. Moreover, the approaches applied here can be easily adapted in other livestock populations.

5.1 Genotyping methods

Various sources of genotypic data were available for analysis. Medium density (~ 50’000) array-derived variants were available for both dam and sire line SLW animals. By analysing the pedigree of these populations, 35 key ancestors were selected from each population, for which high-coverage were generated sequences. Low-pass sequences were available for another dam line.

Microarrays offer high-quality genotype data, have a low cost per sample, and provide accurate genotyping for common or specific variants. However, they are limited in their ability to identify de novo variants, as they rely on pre-defined set of variants. Benefits of WGS include generation of high-quality genotypes, discovery of novel variants, increased power to detect rare variants, and high accuracy in genome-wide

genotyping [1, 2]. However, there are certain disadvantages associated with WGS. First, it has a high cost per sample, making large-scale studies financially challenging. Consequently, a lower number of analyzed samples limits its application in studies requiring a large cohort. Therefore, careful selection of the target animals is desirable. Secondly, data analysis for WGS can be complex due to the vast amount of information generated [3].

Genotyping through low-pass sequencing has several advantages over WGS, as it is affordable and allows for high sample throughput. It provides genotypes for both common and low-frequency variants and demonstrates high genome-wide imputation accuracy for minor allele frequency (MAF) greater than 1% with a minimum coverage of 1x [4]. Furthermore, it can be used as a low-cost and low-accuracy variant discovery tool, making it an accessible option for genetic studies [5]. However, there are some disadvantages associated with low-pass genotyping. The imputation accuracy of rare variants is typically low, and genotyping accuracy is also compromised in difficult regions [6]. Additionally, genotype calls are limited to variants present in the imputation reference panel, which may restrict the scope of analysis [4]. Another drawback demonstrated in Chapter 2 is that heterozygous loci may be called as homozygous for the reference allele («undercalling»). This phenomenon was noticeable from the diagonal elements of the genomic relationship matrix (Figure 2.5). Imputation is expected to improve the usefulness of low-coverage data for the analysis of dominance variation by increasing the amount of genotypic information [7].

5.2 Generation of variant catalogues

The WGS variants used in the chapters 2-4 were called and imputed using different tools. Each variant calling approach has its own advantages and disadvantages, which must be considered when designing the study. Equally, the imputation tools and strategies varied in computational performance and the resulting imputation accuracy. These variant calling approaches allowed for comprehensive analysis of genetic variation across different sample sets and read depths.

5.2.1 Variant calling approaches

Variant calling tools play a crucial role in genomic analysis as they identify and genotype genetic variations from raw sequencing data. Two widely used variant calling tools are GATK (Genome Analysis Toolkit) and DeepVariant. GATK is a comprehensive software package that provides a suite of tools for variant discovery and genotyping. It employs algorithms and statistical models to accurately detect single nucleotide polymorphisms (SNPs), short insertions, and deletions. GATK incorporates best practices for variant calling and follows a robust workflow that includes data preprocessing, variant discovery, and variant quality score recalibration [8]. On the other hand, DeepVariant utilizes deep learning algorithms to call variants from sequencing data. It employs convolutional neural networks to make highly accurate variant calls by modeling the sequencing data patterns and capturing subtle variations. DeepVariant has demonstrated impressive performance in terms of sensitivity and precision, particularly for complex genomic regions [9].

GATK is a widely used tool that has been extensively applied for identifying and genotyping sequence variants in diverse livestock

populations [10, 11]. The accuracy of variant calling is affected by sequence quality, uniformity of coverage, and the threshold of false-discovery rate that is used. Recent studies have suggested that DeepVariant exhibits superior accuracy in genotyping compared to GATK [12, 13]. However, it is important to note that DeepVariant has been less frequently employed for variant calling in species other than humans.

In this thesis, variant calling was performed using both tools. GATK was used in Chapters 2 and 3 to call SNPs, and short insertions, and deletions from high-coverage WGS data and from 421 samples with mixed coverages, respectively. DeepVariant was used in Chapter 4 with high-coverage data. The number of genotyped SNPs was comparable between Chapter 2 and Chapter 4 despite their similar read depths and Chapter 4 containing twice the number of samples. This suggests that the 70 key ancestors sequenced in Chapter 2 already capture a large proportion of the genetic variation present in the two populations. Consequently, the inclusion of additional animals in Chapter 4 did not result in a significant increase in new genetic variants being discovered.

5.2.2 Imputation

Genotype imputation is a pivotal process in genomic studies, enabling the estimation of missing genotypes and the inference of unobserved variants. Various imputation approaches were used in this thesis to enhance the genotyping accuracy and completeness of the datasets.

In Chapter 2, Beagle v.4.1[14] was used for refining and imputation, leveraging the high-coverage sequences of 70 key ancestors to impute almost 27 million short variants, including SNPs, insertions, and deletions. This informative haplotype reference panel facilitated both subsequent

imputations presented in Chapter 2 and Chapter 3. Specifically, in Chapter 2, genotypes at 23 million biallelic SNPs were inferred using GLIMPSE1 [4] method in 175 low-coverage samples (on average 1-fold), and in Chapter 3 SHAPEIT4 tool [15] was used to pre-phase the reference panel followed by Beagle v.5.2 [16] for direct imputation of medium-density array-derived genotypes from 14 thousand samples to the WGS level, leveraging 22 million SNPs with MAF above 1%. The proportion of correctly imputed genotypes, i.e., concordance ratio, was equally 97% for both GLIMPSE and SHAPEIT4 + Beagle v.5.2, and lower (85%) for GATK + Beagle v.4.1. Earlier studies have also found superior imputation accuracy of GLIMPSE over Beagle v.4.1 for low-coverage data and of phasing performance over SHAPEIT4 [4].

Both imputation approaches rely on haplotype phasing, which involves assigning genotypes to their two paternal origins. The accuracy of imputation is greatly influenced by the size and diversity of the reference panel used. A larger and more diverse reference panel increases the chances of capturing a wide range of genetic variation, leading to improved imputation accuracy [17–19]. This is particularly important for rare (MAF < 0.5%) and low-frequency (0.5% < MAF < 5%) variants. Large reference panels have the capability to accurately impute variants with frequencies as low as 0.1–0.5% and can already encompass thousands of putative loss-of-function alleles [17]. This is especially true if the reference panels include animals from the target population [20]. The reference panels in Chapters 2 and 3 were based on samples from the target population. The inclusion of key ancestors, which captured the majority of the genetic diversity in both lines, led to strong LD between imputed and reference variants and consequently to high imputation accuracy.

5.3 Dam and sire lines are diverged

The WGS data from the key ancestors were directly employed in Chapter 2 to assess the separation of the dam and sire lines. Weir and Cockerham F_{ST} values above 0.25 throughout the genome (Supplementary Fig. S2.2) and the separation of the two SLW lines by the top principal components (Supplementary Fig. S2.4) showed that the populations diverged. Although the exchange of genes between the lines stopped only few generations ago, specifically the use of dam line animals in breeding population of the sire line, all samples were separated into clusters. The principal components plot in Chapter 3 (Additional file 3.10) was based on SLW animals that were genotyped with arrays during routine screening by the breeding company. This plot also included animals born before 2018, which was different from Chapter 2. It showed that some animals did not fully belong to one of the lines; however, these animals were mostly labeled as dam line animals. Most of the admixed animals were born before the year 2000, therefore before the separation of the lines (Figure 5.1). Admixture analyses that considered genes from key ancestors revealed only one admixed sample, which had a maximum proportion of 80:20 (Supplementary Fig. S2.1).

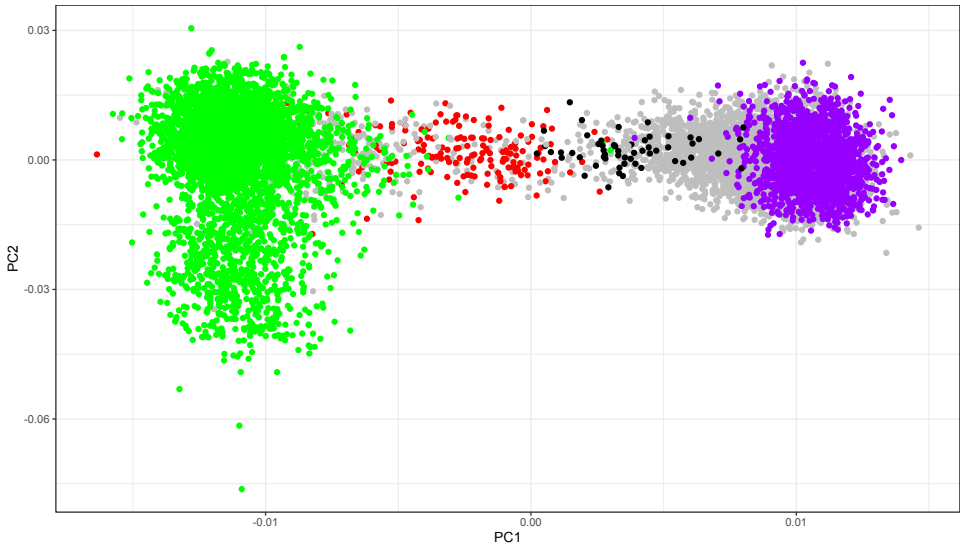


Figure 5.1. Principal component analyses plot (extension of Additional file 3.10). Animals are categorized based on birth year: in black and red are dam and sire line animals born before 2001, and in purple and green are dam and sire line animals born after 2015, respectively. In grey are all others, i.e., born in years 2001-2015.

The distance between clusters is relative; including animals from genetically different breeds reduces the distance between SLW populations. Plotting them together with genetically more distant breeds would scale the principal components and the two lines would cluster much closer. To demonstrate this, I randomly selected 100 animals from each of four breeds (the dam and sire lines from SLW, Pietrain, and Duroc), which had array-derived genotypes (Figure 5.2). The principal components were calculated as described in Chapter 2. Only genotypes with a frequency above 0.5% were used. The first and second principal components effectively separated the breeds into three clearly distinguishable clusters. Additionally, the analysis brought the two lines closer together, resulting

in partial overlap between them. However, it is important to note that the lines still exhibit some separation and are not entirely overlapping.

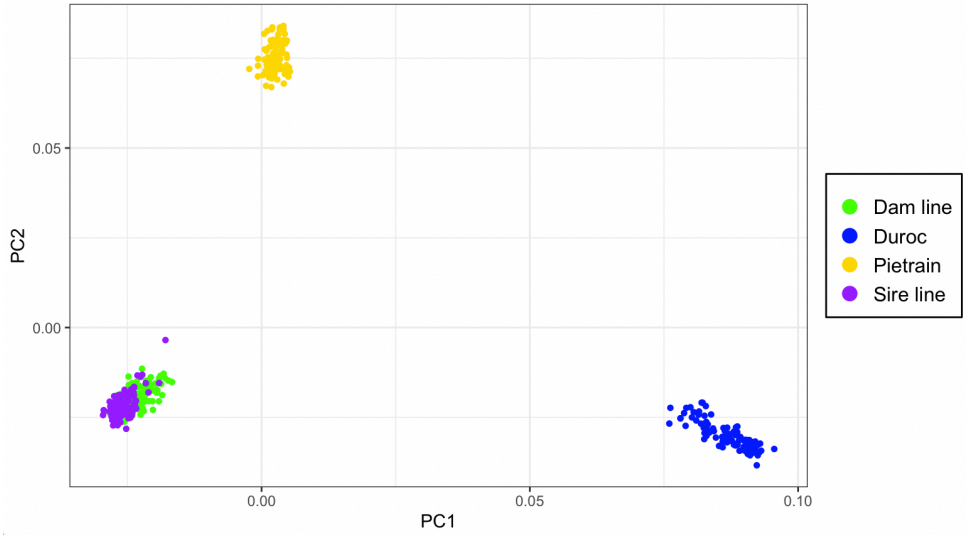


Figure 5.2. Principal component analyses plot of four breeds. Analyses of 100 randomly sampled animals from each breed puts in perspective the separation of the dam and sire lines.

5.4 Consequences of small effective population size

5.4.1 Effective population size is smaller in sire line

The Swiss Large White breed was originally one population, from which the sire line was gradually bred. Creation of the sire line resulted in population reduction, which implied a change in the effective population size [21]. Indeed, effective population sizes (estimated based on pedigree-derived inbreeding coefficients) of both lines were relatively small, especially in the sire line with 44 individuals. This size was smaller than in other pig breeds [22–26] and other livestock species [22, 27, 28]. In chicken lines divergently selected for high and low body weight, Márquez et al. [29]

found lower effective population sizes (~ 35) after long-term divergent breeding and concluded that this is expected in closed lines. The estimation of effective population size is expected to be close to the true value if the pedigree is complete [30], which was the case for this study. Lower effective population size of the sire line may indicate that some boars were more widely used than others, likely due to artificial insemination [29].

5.4.2 Inbreeding levels are higher in sire line

The effective population size is a useful measure to evaluate also the inbreeding levels in animal populations [31]. Thus, a few animals, selected based on their marginal genetic contribution to the active breeding population, account for a majority of the population's haplotype diversity. The sire line, which is characterized by a smaller effective population size, has fewer segregating haplotypes within the population, which in turn leads to higher levels of inbreeding. As observed in Chapter 2, the results corroborated the higher inbreeding level of the sire line compared to the dam line.

Inbreeding was evaluated using two approaches: runs of homozygosity (F_{ROH}) and the genetic relationship matrix derived from the pedigree (F_{PED}). The F_{PED} values were 0.05 ± 0.01 for the dam line and 0.07 ± 0.02 for the sire line, which were lower than the values estimated through runs of homozygosity (F_{ROH}, 0.26 ± 0.03 for the dam line and 0.28 ± 0.03 for the sire line). Indeed, the sire line exhibited a significantly higher proportion of long runs of homozygosity compared to the dam line (Figure 2.3). This finding aligns with the shorter time for the development of the sire line and its smaller effective population size.

A moderate Pearson's correlation coefficient between the F_{PED} and F_{ROH} estimates (r = 0.34) is consistent with the results of a study on Large White by Shi et al. [23]. It is important to note that the F_{PED} estimate represents the statistical expectation of the probable genomic proportion of alleles that are identical by descent (IBD). However, it does not account for the stochastic recombination events that occur during meiosis. As a result, F_{PED} underestimates the true relatedness among individuals [32]. On the other hand, genomic estimates based solely on genetic markers, such as the genomic relationship matrix (F_{GRM}) or homozygosity level (F_{HOM}), do not differentiate between alleles that are identical by state (IBS) and those that are IBD. This can lead to overestimation of inbreeding levels [33]. F_{ROH}, on the other hand, helps mitigate these issues [34].

Fisher [35] reported that the expected length of a DNA segment that is IBD follows an exponential distribution with a mean derived from centimorgan, which represents the average number of recombination events occurring in one generation. Recombination events have the potential to disrupt long chromosome segments, so the presence of long runs of homozygosity indicates recent inbreeding, while short runs of homozygosity are indicative of more distant ancestors. Therefore, the higher presence of long runs of homozygosity in the sire line suggests a greater occurrence of recent inbreeding compared to the dam line.

5.5 Genome-wide association studies and fine-mapping of multi-trait QTL

5.5.1 Use of LD in imputation and GWAS

The LD between causative variants and genetic markers is a fundamental assumption in genomic studies investigating the relationship

between phenotypes and genotypes. Correspondingly, the accuracy of standard imputation techniques is strongly affected by the similarity of LD patterns in the target and reference populations [36].

In Chapter 3, the imputation proved to be rather accurate due to the use of a reference panel comprised of animals from the same population as the target population (Additional file 3.10). Despite the limited availability of high-density array-derived genotypes, which resulted in relatively high LD blocks (Figure 3.2 B), previously identified candidate causal variants were successfully recovered among the top variants in four out of six multi-trait QTL.

5.5.2 Single- and multi-trait GWAS with 24 traits reveal in total 15 QTL

The association analyses in Chapter 3 incorporated 24 economically relevant traits and 40K chip genotypes of 5,753 pigs. Single-trait association analyses revealed eleven QTL, and various settings of the multi-trait or meta-analyses GWAS revealed between three and seven QTL.

There are several reasons that may contribute to the scarcity of GWAS signals. First, the significance threshold influences what is deemed a QTL. As discussed by van den Berg et al. [37], there is a growing need for an appropriate method to adjust for the number of independent tests because of increased marker density, which results in stronger LD and population stratification. In this thesis was used a Bonferroni correction with the total number of SNPs (1.24×10^{-6} and 3.11×10^{-9} for array genotypes and imputed sequence variants, respectively). Becker at al. [38] conducted association analyses on the same population but identified four

QTL with the two traits by defining a moderate-significance threshold at P-value = 5×10^{-5}. The QTL regions identified by Becker et al. [38] were nearly significant but failed to pass the stringent significance threshold.

Secondly, the density of the array genotypes may be insufficient. LD decay between a SNP and the causative mutation, due to recombination events, can weaken the association signal. Genotyping the causative mutation directly or increasing the chances of capturing it would help address this issue.

5.5.3 SNP heritabilities agree with those routinely evaluated

Genetic architecture of the traits plays another important role in success of the GWAS. Genomic heritability measures the amount of variation that would be explained by GWAS when the sample size is so large that all associated variants would be statistically significant [39]. Classical estimation of total narrow-sense heritability (estimated from phenotypic records of samples that include family members) captures the total amount of additive genetic variance in the population irrespectively of the joint distribution of allele frequency and effect size [40]. In contrast, SNP heritability (estimated from genotype data) captures only the proportion of additive genetic variance due to LD between the SNPs and the unknown causal variants.

SNP heritability estimates of the deregressed breeding values of the traits are shown in Table 3.1. To analyze the traits' SNP heritability, a model with the weights as implemented in GCTA was used. The GCTA algorithm for estimating variance components uses average information (AI) methods, i.e., likelihood-based iterated inference procedure. By including weights, the diagonal elements of the residual correlation matrix

in REML were corrected. The resulting SNP heritability estimates were highly correlated (r^2 = 96%) to those obtained from routine evaluation while using raw phenotypes (narrow-sense heritability).

The SNP heritability estimates while using weights were on average 2x lower than when estimating without the weights. Despite the magnitude of this difference, the effect of weights is assumed to be minimal, as shown on breeding value estimation by Wang et al. [41]. Supporting the use of weights, other studies found similar or only slightly different heritability estimates with pedigree-based model in different Large White populations [42–44]. However, Hong et al. [45] found substantially lower SNP-based heritability estimates than reported in the Chapter 3 (for body length by 0.42 and number of teats by 0.19).

5.5.4 Only fraction of the total variance was captured by QTL

The multi-trait QTLs identified in Chapter 3 only accounted for a fraction of the overall variation in the traits (Figure 3.2C). Proportion of the total phenotypic variation explained by alternative alleles of the lead SNPs was on average 6.5 %. Potential lies in rare variants with MAF < 1% [46], structural variants [47], tandem repeats [48, 49], or long non-coding RNA [49]. Particularly reproduction traits, with which I did not reveal any associations in this thesis, have been reported to be affected by structural variation [50]. However, the detection of structural variations spanning longer fragments remains challenging when using array-derived genotypes or short reads [49, 51], they were not considered in this thesis.

5.5.5 Largest effect sizes revealed for carcass length

Cumulatively the most explained total variation was from 17 % for trait from conformation group ‘carcass length’. The two loci were on chromosomes 7 and 17, both with previously proposed causal genes *VRTN* and *BMP2*, respectively. Apart from those two genes, there are other potentially causal, large effect genes segregating in pig populations. Two loci, *PLAG1* and *LCORL*, together explained 18.4% of the residual variance in body length in an intercross between Large White pigs and wild boar [52].

A locus with a large effect on a favorable trait tends to be strongly selected and fixed in domestic animals [53]. For example, in cattle, several major loci explained more than a third of genetic variance [54]. In horses, only the LCORL region explained 11% and 18% of the phenotypic variance in Franches-Montagnes and German Warmblood, respectively [55, 56]. Makvandi-Nejad et al. [57] reported that four loci explained 83% of size variance in the 48 horses from 16 breeds and two loci explained 59% of the variance in thoroughbred size, although these estimates are likely to be upwardly biased by the small and selected sample. In contrast, adult human height explains ~0.3% to ~0.5% of the phenotypic variance [58],

5.6 Functional and structural annotation needs to be improved

It is increasingly recognized that most of the genetic variation responsible for complex traits, such as common complex diseases in humans or economically significant traits in plants and animals, can be attributed to regulatory variants rather than coding variants [59, 60]. In the

Chapter 3 were used imputed both SNPs and short insertions and deletions to reveal the associations. The meta-analyses multi-trait GWAS of 24 traits yielded 9,774 significantly associated variants in 159 genes across six QTL. I annotated these variants using VEP (cache v. 104; [61]) allowing for multiple consequence predictions (Table 5.1). Indeed, vast majority (73 %) of the significantly associated variants were located in the intronic regions, including all the top variants (Chapter 3).

Introns, non-coding segments of DNA or RNA, were once considered as "junk DNA". However, increasing evidence [62–64] highlights their crucial involvement in genomic regulation. Intronic regions play a direct role in alternative splicing, gene expression, mRNA transport, chromatin assembly, and nonsense-mediated decay [62, 65]. Moreover, they have indirect implications, such as influencing the functional characteristics of a gene based on its position in the sequence, impacting evolutionary processes, providing a source of new genes, and harboring non-coding functional RNA genes [62].

Table 5.1: Predicted consequence types and counts of 9,774 significantly associated variants in the 6 multi-trait QTL detected by meta-analyses GWAS.

Consequence type (all)	***Number of occurances***
Intron variant	42,484
Upstream gene variant	4,123
Intron variant and non-coding transcript variant	3,597
Downstream gene variant	3,408
Intergenic variant	3,174
3` prime UTR variant	446
Synonymous variant	307
Non-coding transcript exon variant	241
5` prime UTR variant	106

Missense variant	83
Splice region variant and intron variant	79
Frameshift variant	11
Stop gained	11
Frameshift variant and splice region variant	3
Frameshift variant and splice region variant and intron variant	3
Splice region variant and non coding transcript exon variant	2
Splice region variant and synonymous variant	2
Stop gained and frameshift variant	1
Coding sequence variant	1

Genome annotation can be done as large-scale similarity comparisons on the genome sequence, taking note of repeated regions at different scales, and then looking for function in the genome by mapping the 'read-out' from experiments onto sequence elements [66–68]. Thus, separating the analysis into comparative and functional. Annotation can be a manual or automated process [69]. While predicting the consequences of alleles on overlapping transcripts can help narrow down potential causal variants, it is important to note that automated annotation is not always reliable [70, 71]. Automated annotation of large, fragmented “draft” genomes remains difficult; in addition, errors and contamination in draft assemblies lead to errors in annotation that tend to propagate across species.

Human and other vertebrate genome annotations are provided by reputable sources, such as Ensembl or GENCODE. These resources offer annotations for protein-coding genes, long non-coding RNAs (lncRNAs), small non-coding RNAs (sRNAs), pseudogenes of protein-coding genes, as well as immunoglobulin, and T-cell receptor segments. They continually update the annotation by (i) identifying novel protein-coding genes, lncRNAs, and pseudogenes, (ii) capturing newly discovered alternatively

spliced transcripts in protein-coding and lncRNA regions, (iii) iteratively reassessing existing genes and transcripts to ensure accuracy, and (iv) integrating novel biological features into the annotation to enhance its comprehensiveness [72].

Gene transfer files (GTF) contain information about gene structure. The features included in Ensembl GTF file for the Sscrofa11.1 pig assembly (v. 107) are shown in Figure 5.3. Known positions of the start and stop codons of the genes enabled determination the positions of splice sites, intronic, and intergenic regions (Chapter 4). Next, a missing genic feature - branch point sequence (BPS) was annotated in 30 species, including the pig (the 192,744 predicted BPS positions in the pig genome are listed in Supplementary Table S5.1). The annotation approach employed in this thesis included prediction of the positions built upon detection of BPS in human genome and recognition of the motif similarity across species [73].

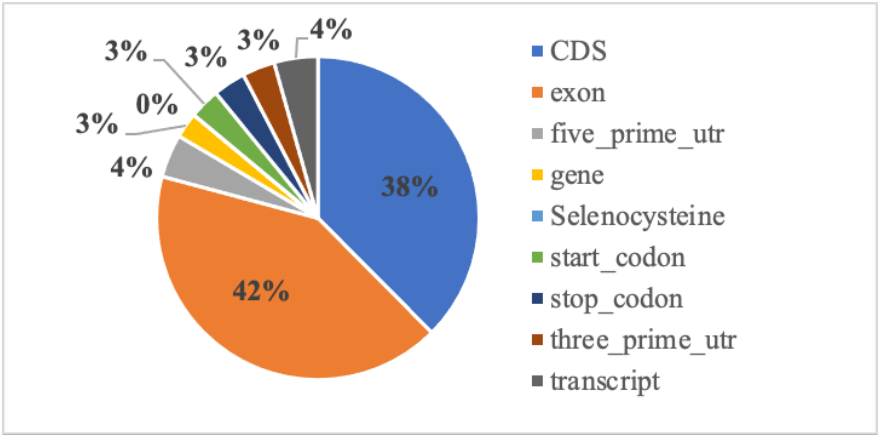


Figure 5.3: Genic features included in gene transfer file from Ensembl for Sscrofa11.1 assembly (v.107).

The prediction of the genic features included in splicing machinery is necessary for a correct assessment of the consequences of variants located in these sequences. Another intronic feature not included in the annotation file is the polypyrimidine tract. These poly-A stretches, located in proximity of 3' end of the intron, promote the assembly of the spliceosome [74]. The sequence between the branch point and the 3' splice site, which encompass the polypyrimidine tract, is devoid of AG dinucleotides, creating a so-called 'AG exclusion zone' [75]. The mutations creating AG result in mis-splicing events with pathogenic consequences [76, 77].

Many prediction tools of various features are being currently developed [67–69, 78–82]. On the other hand, these approaches yield some false positives and overlook sequences located in exceptional positions, i.e., distant branch points further from the splice site than the majority [83]. Therefore, experimental methods are therefore still needed for verification of the computational predictions [84]. The false positive predicted BPS in each intron were reduced by including only the predictions with the highest score. On the other hand, this approach might remove some of the true BPS.

Intergenic regions also harbor features affecting the gene regulation. To facilitate the identification of regulatory variants and to assess their impact on gene expression, the animal genomics community has initiated the creation of epigenome maps [82]. These maps primarily rely on techniques such as ChIP-Seq [85] (chromatin immunoprecipitation followed by sequencing), DNase-Seq [86] (DNase I hypersensitive site sequencing), and ATAC-Seq [87] (assay for transposase-accessible chromatin using sequencing). The international Functional Annotation of Animal Genomes (FAANG) project plays a key role in coordinating most of these initiatives [82]. Liver-specific comparative enhancer maps,

utilizing histone modification data, have already been constructed for 20 mammals, including pig, cow, and rabbit [88].

The findings from Chapter 3 and other association studies indicate that most genetic variants associated with disease susceptibility are not situated in protein-coding regions [89, 90]. Instead, over 90% of disease-associated, non-coding SNPs in humans are located within regulatory elements [91, 92]. Thus, improvement of the genome annotation is crucial for uncovering the new causal variants of both quantitative and qualitative traits.

5.7 Conclusion and future directions

This thesis aimed to resolve major genetic questions related to the two lines of SLW pig breed. To tackle them, a variety of genomic data were generated and considered, including arrays and short read sequences. Imputation enabled analyses with larger sample sizes. The population structure and selection signatures within and between the lines were examined, the single- and multi- trait associations were detected, and positions of a genic feature for enhancement of the current annotation were predicted.

Economically relevant traits in livestock can be in linkage disequilibrium with deleterious alleles that cause recessive disorders. One approach to detect the association is to use proxy-phenotypes, such as lactation or growth and developmental traits, and GWAS in non-additive inheritance mode. For example, this approach was successfully applied in a large cohort of New Zealand dairy cattle [93]. It yielded five novel QTL

associated with large recessive impacts on three milk yield traits. This method can be implemented in any breeding population because routinely collected phenotypes can be used as proxy for rare disorders, for which the data are often missing. In this thesis, direct phenotypes were not utilized, but they could be easily incorporated into the GWAS approach used here. Additionally, the exploration of changing to a non-additive inheritance mode is a possibility and could be worthwhile to investigate.

Further research can be conducted to improve annotation. Chapter 4 of this thesis closely examined the branch point sequences. Other genomic features, such as DNA repeats or methylation pattern, are of interest due to next layer influence on the inheritance and appearance of the traits. For example, short tandem repeats in swine genome have been associated with breed identity at a greater extent than SNPs [48]. Tackling these topics requires generation of new data. Specifically, this requires long read sequencing to detect structural variation, RNA-seq to precisely quantify the transcriptome and study allele-specific expression, or ChIP-Seq to characterize the epigenome. An improved annotation can help uncover the underlying genetic mechanism behind some traits. Fine-mapping with inclusion of functional annotations as well as transcriptomic data can be done, for example, in GPA-MDS – multivariate GWAS simultaneously modeling for annotation [94].

The genomic variants analyzed in this thesis focused on SNPs and short insertions and deletions located on the 18 porcine autosomes. However, the inclusion of other types of genomic variation may uncover new associations. For example, chromosomal abnormalities, and

particularly structural chromosome rearrangements, that are remarkably prevalent in the domestic pig relative to other species [95]. There are reports describing chromosome abnormalities associated with physical malformations in SLW [96] or lower litter sizes [97, 98]. The litter size losses caused by chromosome rearrangements are variable among carriers and are dependent on a variety of factors, including the morphology of the rearrangement [98]. Another example of chromosomal rearrangements are chromosomal aneuploidies. These numerical chromosomal aberrations are usually lethal when occurring in autosomes, hence aneuploidy is rarely seen in living individuals, except for aneuploidy in sex chromosomes and down syndrome in humans [99]. Although it is a rare phenomenon in liveborn individuals, it is observed in livestock breeding populations [100]. Preliminary analyses have shown that the chromosome rearrangements might be heritable [101].

SLW populations are routinely genotyped, and the genotype intensity data from SNP arrays can be utilized to identify aneuploidy [102, 103]. However, this identification process is time-consuming and costly, typically involving visual inspection of the data per chromosome through plots of intensity data by an expert. One approach to streamline this process is to incorporate the identification of SNPs closely linked to chromosome rearrangements into cytogenomics analyses. This can serve as a control effort to identify boars at risk of producing carrier offspring. Alternatively, a recent development by Bouwman et al. [100] offers a more efficient solution. They developed a deep learning Convolutional Neural Network (CNN) classification model that operates based on chromosome-level plots of SNP array intensity data. This innovative model can accurately classify images into disomic, monosomic, and trisomic cases. By employing this

classification model, routine screening becomes more effective. These approaches have the potential to improve efficiency and accuracy in identifying aneuploidy cases, ultimately benefiting breeding programs and the overall management of SLW populations.

5.8 References

1. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen JB, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Nephrol. 2014;15:1–15.
2. Chat V, Ferguson R, Morales L, Kirchhoff T. Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies. Front Genet. 2022;12:2712.
3. Pfeifer SP. From next-generation resequencing reads to a high-quality variant data set. Heredity (Edinb). 2016;118:111–24.
4. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53:120–6.
5. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera A V. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. Genome Med. 2019;11:1–12.
6. Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, et al. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. 2021;22:197.
7. Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genetics Selection Evolution. 2015;47:1–14.
8. Technical Documentation – GATK. https://gatk.broadinstitute.org/hc/en-us/categories/360002310591. Accessed 1 Jun 2023.
9. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.

10. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, et al. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science (1979). 2009;324:522–8.

11. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science (1979). 2009;323:133–8.

12. Lloret-Villas A, Bhati M, Kumar Kadri N, Fries R, Pausch H. Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. bioRxiv. 2021;:1–24.

13. Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, et al. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nat Commun. 2020;11:1–14.

14. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.

15. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019;10:1–10.

16. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103:338–48.

17. Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet. 2015;23:975–83.

18. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.

19. Al Bkhetan Z, Zobel J, Kowalczyk A, Verspoor K, Goudey B. Exploring effective approaches for haplotype block phasing. BMC Bioinformatics. 2019;20:540.

20. Lloret-Villas A, Pausch H, Leonard AS. The size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle. Genet Sel Evol. 2023;55.

21. Ramos-Onsins SE, Burgos-Paz W, Manunza A, Amills M. Mining the pig genome to investigate the domestication process. Heredity (Edinb). 2014;113:471–84.

22. Hall SJG. Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. Animal. 2016;10:1778–85.

23. Shi L, Wang L, Liu J, Deng T, Yan H, Zhang L, et al. Estimation of inbreeding and identification of regions under heavy selection based on runs of homozygosity in a Large White pig population. J Anim Sci Biotechnol. 2020;11:1–10.

24. Frantz LAF, Madsen O, Megens H-J, Bosse M, Paudel Y, Crooijmans RPMA, et al. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biol. 2013;14.

25. Zanella R, Peixoto JO, Cardoso FF, Cardoso LL, Biegelmeyer P, Cantão ME, et al. Genetic diversity analysis of two commercial breeds of pigs using genomic and pedigree data. Genet Sel Evol. 2016;48:24.

26. Grossi DA, Jafarikia M, Brito LF, Buzanskas ME, Sargolzaei M, Schenkel FS. Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. BMC Genet. 2017;18.

27. Liu S, Reddy UK, Fereň cakoví M, Miar Y, Karimi K, Hossain Farid A, et al. Linkage Disequilibrium, Effective Population Size and Genomic Inbreeding Rates in American Mink Using Genotyping-by-Sequencing Data. Front Genet. 2020;11.

28. Rodríguez-Ramilo ST, Fernández J, Toro MA, Hernández D, Villanueva B. Genome-Wide Estimates of Coancestry, Inbreeding and Effective Population Size in the Spanish Holstein Population. PLoS One. 2015;10:e0124157.

29. Márquez GC, Siegel PB, Lewis RM. Genetic diversity and population structure in lines of chickens divergently selected for high and low 8-week body weight. Poult Sci. 2010;89:2580–8.

30. Boichard D, Maignel L, Verrier E. The value of using probabilities of gene origin to measure genetic variability in a population. Genet Sel Evol. 1997;29.

31. Theodorou K, Couvet D. On the expected relationship between inbreeding, fitness, and extinction. Genet Sel Evol. 2006;38.

32. Henryon M, Liu H, Berg P, Su G, Nielsen HM, Gebregiwergis GT, et al. Pedigree relationships to control inbreeding in optimum-contribution

selection realise more genetic gain than genomic relationships. Genetics Selection Evolution. 2019;51:1–12.

33. Wang J. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. J Evol Biol. 2014;27:518–30.

34. Peripolli E, Stafuzza NB, Munari DP, Lima ALF, Irgang R, Machado MA, et al. Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. BMC Genomics. 2018;19.

35. Fisher RA. A fuller theory of "Junctions" in inbreeding. Heredity (Edinb). 1954;8:187–97.

36. Kabisch M, Hamann U, Lorenzo Bermejo J. Imputation of missing genotypes within LD-blocks relying on the basic coalescent and beyond: Consideration of population growth and structure. BMC Genomics. 2017;18:1–12.

37. van den Berg S, Vandenplas J, van Eeuwijk FA, Lopes MS, Veerkamp RF. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. Journal of Animal Breeding and Genetics. 2019;136:418–29.

38. Becker D, Wimmers K, Luther H, Hofer A, Leeb T, Moore S. A Genome-Wide Association Study to Detect QTL for Commercially Important Traits in Swiss Large White Boars. PLoS One. 2013;8.

39. Vinkhuyzen AAE, Wray NR, Yang J, Goddard ME, Visscher PM. Estimation and Partitioning of Heritability in Human Populations using Whole Genome Analysis Methods. Annu Rev Genet. 2013;47:75.

40. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008;9:255–66.

41. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb). 2012;94:73–83.

42. Dong L, Tan C, Cai G, Li Y, Wu D, Wu Z. Estimates of variance components and heritability using different animal models for growth, Backfat, litter size, and healthy birth ratio in Large White Pigs. Can J Anim Sci. 2020;100:330–6.

43. Krupa E, Wolf J. Simultaneous estimation of genetic parameters for production and litter size traits in Czech Large White and Czech Landrace pigs. Czech J Anim Sci. 58:429–36.

44. Ogawa S, Konta A, Kimata M, Ishii K, Uemoto Y, Satoh M. Estimation of genetic parameters for farrowing traits in purebred Landrace and Large White pigs. Anim Sci J. 2019;90:23.

45. Hong Y, Ye J, Dong L, Li Y, Yan L, Cai G, et al. Genome-Wide Association Study for Body Length, Body Height, and Total Teat Number in Large White Pigs. Front Genet. 2021;12:1218.

46. Wainschtein P, Jain D, Zheng Z, Aslibekyan S, Becker D, Bi W, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat Genet. 2022;:1–11.

47. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610:704–12.

48. Wu Z, Gong H, Zhang M, Tong X, Ai H, Xiao S, et al. A worldwide map of swine short tandem repeats and their associations with evolutionary and environmental adaptations. Genet Sel Evol. 2021;53:39.

49. Du H, Zheng X, Zhao Q, Hu Z, Wang H, Zhou L, et al. Analysis of Structural Variants Reveal Novel Selective Regions in the Genome of Meishan Pigs by Whole Genome Sequencing. Front Genet. 2021;12:550676.

50. Mo J, Lu Y, Zhu S, Feng L, Qi W, Chen X, et al. Genome-Wide Association Studies, Runs of Homozygosity Analysis, and Copy Number Variation Detection to Identify Reproduction-Related Genes in Bama Xiang Pigs. Front Vet Sci. 2022;9:892815.

51. Jiang YF, Wang S, Wang CL, Xu RH, Wang WW, Jiang Y, et al. Pangenome obtained by long-read sequencing of 11 genomes reveal hidden functional structural variants in pigs. iScience. 2023;26.

52. Rubin C-JJ, Megens H-JJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109:19529–36.

53. Takasuga A. PLAG1 and NCAPG-LCORL in livestock. Animal Science Journal. 2016;87:159–67.

54. Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics. 2014;15.

55. Tetens J, Widmann P, Kühn C, Thaller G. A genome-wide association study indicates LCORL/NCAPG as a candidate locus for withers height in German Warmblood horses. Anim Genet. 2013;44:467–71.

56. Signer-Hasler H, Flury C, Haase B, Burger D, Simianer H, Leeb T, et al. A genome-wide association study reveals loci influencing height and other conformation traits in horses. PLoS One. 2012;7.

57. Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, Gu E, Chandler AM, et al. Four loci explain 83% of size variation in the horse. PLoS One. 2012;7.

58. Visscher PM. Sizing up human height variation. Nat Genet. 2008;40:489–90.

59. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.

60. Huang H, Fang M, Jostins L, Umićević Mirkov M, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;547:173–8.

61. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17.

62. Jo B-S, Choi SS. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. 2015;13:112.

63. Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. 2018;32:577–91.

64. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338:1593–9.

65. Keller EB, Noon WA. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc Natl Acad Sci U S A. 1984;81:7417.

66. García-Sancho M, Lowe J. A History of Genomics across Species, Communities and Projects. 2023. https://doi.org/10.1007/978-3-031-06130-1.

67. König S, Romoth L, Stanke M. Comparative Genome Annotation. Methods Mol Biol. 2018;1704:189–212.

68. Armstrong J, Fiddes IT, Diekhans M, Paten B. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci. 2019;7:41–64.

69. García-Sancho M, Lowe J. Making Reference Genomes Useful: Annotation. 2023;:205–54.

70. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47:10994–1006.

71. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 2019;20:1–3.

72. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.

73. Zhang Q, Fan X, Wang Y, Sun M-A, Shao J, Guo D. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics. 2017;33:3166–72.

74. Sharma S, Kohlstaedt LA, Damianov A, Rio DC, Black DL. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat Struct Mol Biol. 2008;15:183–91.

75. Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, Smith CWJ. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. Genome Biol. 2006;7:1–19.

76. Bryen SJ, Yuen M, Joshi H, Dawes R, Zhang K, Lu JK, et al. Prevalence, parameters, and pathogenic mechanisms for splice-altering acceptor variants that disrupt the AG exclusion zone. Human Genetics and Genomics Advances. 2022;3:100125.

77. Wimmer K, Schamschula E, Wernstedt A, Traunfellner P, Amberger A, Zschocke J, et al. AG-exclusion zone revisited: Lessons to learn from 91 intronic NF1 3′ splice site mutations outside the canonical AG-dinucleotides. Hum Mutat. 2020;41:1145.

78. Signal B, Gloss BS, Dinger ME, Mercer TR. Machine learning annotation of human branchpoints. Bioinformatics. 2018;34:920–7.

79. Brůna T, Li H, Guhlin J, Honsel D, Herbold S, Stanke M, et al. GALBA: Genome Annotation with Miniprot and AUGUSTUS. 2023. https://doi.org/10.1101/2023.04.10.536199.

80. Karlsson M, Sjöstedt E, Oksvold P, Sivertsson Å, Huang J, Álvez MB, et al. Genome-wide annotation of protein-coding genes in pig. BMC Biol. 2022;20:1–18.

81. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38.

82. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16.

83. Královičová J, Borovská I, Pengelly R, Lee E, Abaffy P, Šindelka R, et al. Restriction of an intron size en route to endothermy. Nucleic Acids Res. 2021;49:2460–87.

84. Liu Y, Nie H, Liu H, Lu F. Poly(A) inclusive RNA isoform sequencing (PAIso−seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. Nat Commun. 2019;10:1–13.

85. Chromatin Immunoprecipitation Sequencing (ChIP-Seq). https://emea.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html. Accessed 24 Jul 2023.

86. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;2010.

87. ATAC-Seq Analysis of Chromatin Accessibility. https://emea.illumina.com/techniques/popular-applications/epigenetics/atac-seq-chromatin-accessibility.html. Accessed 24 Jul 2023.

88. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160:554–66.

89. French JD, Edwards SL. The Role of Noncoding Variants in Heritable Disease. Trends Genet. 2020;36:880–91.

90. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.

91. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

92. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science (1979). 2012;337:1190–5.

93. Reynolds EGM, Lopdell T, Wang Y, Tiplady KM, Harland CS, Johnson TJJ, et al. Non-additive QTL mapping of lactation traits in 124,000 cattle reveals novel recessive loci. Genet Sel Evol. 2022;54.

94. Wei W, Ramos PS, Hunt KJ, Wolf BJ, Hardiman G, Chung D. GPA-MDS: A Visualization Approach to Investigate Genetic Architecture among Phenotypes Using GWAS Results. Int J Genomics. 2016;2016.

95. Raudsepp T, Chowdhary BP. Cytogenetics and chromosome maps. The Genetics of the Pig: Second Edition. 2011;:134–78.

96. Grahofer A, Letko A, Häfliger IM, Jagannathan V, Ducos A, Richard O, et al. Chromosomal imbalance in pigs showing a syndromic form of cleft palate. BMC Genomics. 2019;20.

97. Pinton A, Ducos A, Berland H, Seguela A, Brun-Baronnat C, Darré A, et al. Chromosomal Abnormalities in Hypoprolific Boars. Hereditas. 2000;132:55–62.

98. Quach AT, Revay T, Villagomez DAF, Macedo MP, Sullivan A, Maignel L, et al. Prevalence and consequences of chromosomal abnormalities in Canadian commercial swine herds. Genetics Selection Evolution. 2016;48:1–7.

99. Hassold T, Hunt P. To err (meiotically) is human: the genesis of human aneuploidy. Nat Rev Genet. 2001;2:280–91.

100. Bouwman AC, Hulsegge I, Hawken RJ, Henshall JM, Veerkamp RF, Schokker D, et al. Classifying aneuploidy in genotype intensity data using deep learning. Journal of Animal Breeding and Genetics. 2023;140:304–15.

101. Donaldson B, Villagomez DAF, King WA. Classical, Molecular, and Genomic Cytogenetics of the Pig, a Clinical Perspective. Animals. 2021;11:1257.

102. Xiong B, Tan K, Tan YQ, Gong F, Zhang SP, Lu CF, et al. Using SNP array to identify aneuploidy and segmental imbalance in translocation carriers. 2014. https://doi.org/10.1016/j.gdata.2014.05.007.

103. Berry DP, O'brien A, O'donovan J, Mchugh N, Wall E, Randles S, et al. Aneuploidy in dizygotic twin sheep detected using genome-wide single nucleotide polymorphism data from two commonly used commercial vendors. 2018. https://doi.org/10.1017/S1751731118000204.

5.9 Supplementary files

Supplementary Table S5.1: Predicted branch point positions.

The Supplementary Table S5.1 is available at https://doi.org/10.5281/zenodo.8202363.
The table includes 192,744 branch point sequence positions predicted from the Susscrofa11.1 assembly. The table lists the gene name, strand, transcript, intron number, chromosome, intron start and end positions, branch point sequence (heptamer with the branch point at position 6), distance from the 3' splice site, and prediction score from the BPP tool.

Adéla (Nosková) Poublan-Couzardot

SPINNEREIWEG 19, 8097 EFFRETIKON, SWITZERLAND

+41 787 52 68 90 • ANOSKOVA@ETHZ.CH

EDUCATION

PhD in Animal Genomics 2023

ETH Zürich, Switzerland

Dissertation title: "Identification of genetic characteristics and trait-associated variants in Swiss pigs"

Committee: Prof. Dr. Hubert Pausch, Prof. Dr. Christine Baes

MSc in Biotechnology and Animal Breeding 2019

Czech University of Life Sciences Prague, The Czech Republic

Thesis title: "Analysis of relationship between conformation and production traits at the level of measured (phenotypic) values "

BSc in Livestock Production 2016

Czech University of Life Sciences Prague, The Czech Republic

Thesis title: „Genetic analysis of milk production "

RESEARCH EXPERIENCE

Scientific Assistant 2019 - 2023

ETH Zürich, Switzerland

Animal Genomics group

Research project: "Whole-genome sequence-based domestic pig breeding "

Research assistant 2018 - 2019

Institute of Animal Science Prague, The Czech Republic

Department of Genetics and Livestock Breeding

Research project: "Improving reliability of national genomic evaluation of dairy cattle"

Erasmus exchange program Feb - Jul 2018

Wageningen University & Research, The Netherlands

Selected courses: Advanced Statistics, Data Management, Modern Statistics for the Life Sciences, Population and Quantitative Genetics

Publications

Nosková, A., Li, Ch., Wang, X., Leonard, A.S., Pausch, H., Kadri, N.K. Exploiting public databases of genomic variation to quantify evolutionary constraint on the branch point sequence in 30 plant and animal species. In submission. 2023.

Nosková, A., Mehrotra, A., Kadri, N.K., Lloret-Villas, A., Neuenschwander, S., Hofer, A., Pausch, H. Comparison of two multi-trait association testing methods and sequence-based fine mapping of six QTL in Swiss Large White pigs. BMC Genomics, 24(192). 2023. https://doi.org/10.1186/s12864-023-09295-4

Nosková, A., Bhati, M., Kadri, N.K., Neuenschwander, S., Hofer, A., Pausch, H. Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. BMC Genomics, 22(290). 2021. Doi: 10.1186/s12864-021-07610-5.

Nosková, A., Hiltpold, M., Janett, F., Echtermann, T., Fang, Z.H., Sidler, X., Selige, C., Hofer, A., Neuenschwander, S. and Pausch, H. Infertility due to defective sperm flagella caused by an intronic deletion in DNAH17 that perturbs splicing. Genetics, 217(2). 2021. Doi: 10.1093/genetics/iyaa033.

Nosková, A., C. Wurmser, D. Crysnanto, A. Sironen, P. Uimari, R. Fries, M. Andersson, and H. Pausch. Deletion of porcine BOLL is associated with defective acrosomes and subfertility in Yorkshire boars. Animal Genetics. 2020. Doi: 10.1111/age.12998.

Fang, Z.H., **Nosková, A.**, Crysnanto, D., Neuenschwander, S., Vögeli, P. and Pausch, H. A 63-bp insertion in exon 2 of the porcine KIF21A gene is associated with arthrogryposis multiplex congenita. Animal Genetics. 2020. Doi: 10.1111/age.12984.

Conference Presentations

Oral Presentations

WCGALP, Rotterdam, The Netherlands 2022

Title: "Detecting QTL for two lowly correlated traits using multi-trait and meta-analysis approaches in Swiss Large White pigs"

In section: "Statistical genetics: GWAS (2) "

EAAP, Davos, Switzerland 2021

Title: "En route to routine genotyping by low-pass sequencing in Swiss pig breeds"

In section: "Exploiting sequence data in breeding programs "

ICAR, Prague, The Czech Republic 2019

Title: "Relationships between conformation traits and milk yield, lifetime production and number of lactations in Czech Holstein cows"

Poster Presentations

IAS Doctoral Symposium, Switzerland 2019

Title: "Whole-genome sequence-based domestic pig breeding"

XXVIIIth Genetic Days, České Budějovice, The Czech Republic 2018

Title: "Analysis of relationship between Conformation and Production Traits at the Level of Measured (phenotypic) Values"

Selected Training

"Genomic prediction considering admixed populations and GxE" Jul 2022

3-day on-site GenTORE Summer School at WUR, The Netherlands.

"Characterization, management and exploitation of genomic diversity in animals" Dec 2019

5-day on-site course held by INRA at WUR, The Netherlands.

Professional experience

Statistician and Data Analyst Sep 2018 - Apr 2019

Median s.r.o., Prague, The Czech Republic

Processing and analysing data in Microsoft SQL server, R studio and local software packages.

Dairy Herd Worker 2016 - 2018

Collective farm, Telč, The Czech Republic