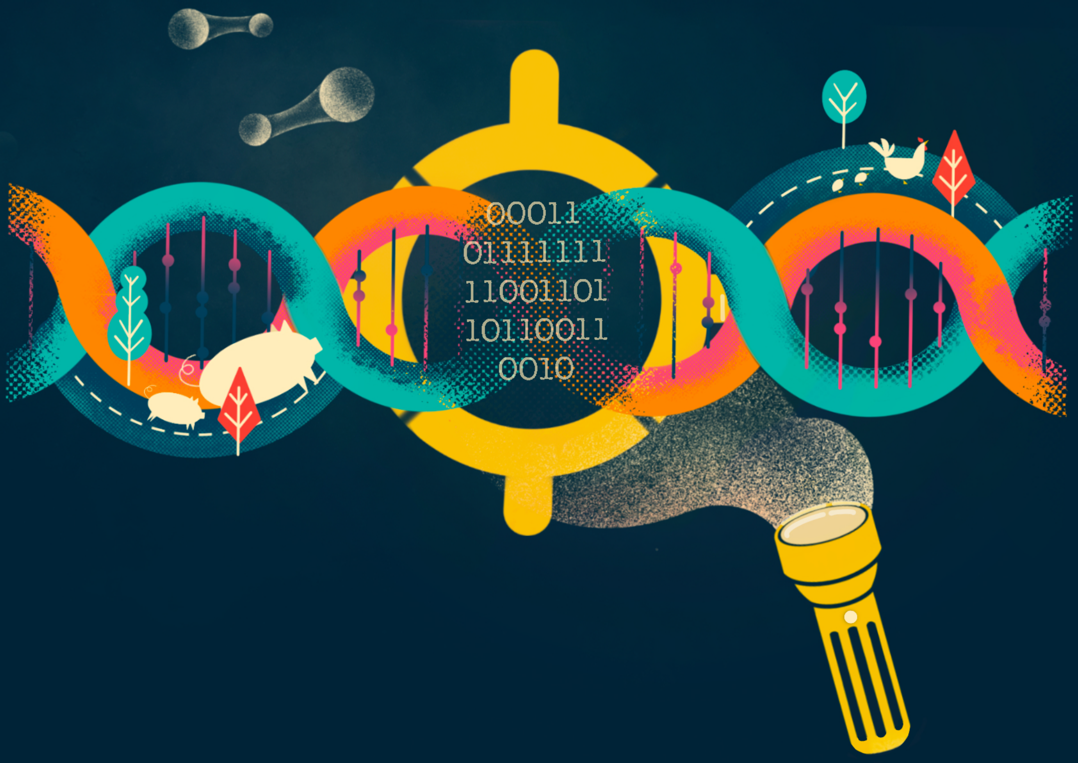
From Sequence to Phenotype:

The Impact of Deleterious Variation in Livestock

Martijn F.L. Derks



From sequence to phenotype: the impact of deleterious variation in livestock

Martijn F.L. Derks

Thesis committee

Promotor
Prof. Dr M.A.M. Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Co-promotor
Dr H.-J.W.C. Megens
Assistant professor, Animal Breeding and Genomics
Wageningen University & Research

Other members
Dr C. Charlier, University of Liège, Belgium
Dr P. Knap, Genus-PIC, Rabenkirchen-Fauluck, Germany
Prof. Dr B.J. Zwaan, Wageningen University & Research, Netherlands
Prof. Dr Cord Drögemüller, University of Bern, Switzerland

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

From sequence to phenotype: the impact of deleterious variation in livestock

Martijn F.L. Derks

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 2 September, 2020
at 4. p.m. in the Aula.

Martijn F.L. Derks
From sequence to phenotype: the impact of deleterious variation in livestock

PhD thesis, Wageningen University, the Netherlands (2020)
With references, with summary in English

ISBN 978-94-6395-320-7
DOI: https://doi.org/10.18174/515207

Abstract

Derks, MFL. (2020). From sequence to phenotype: the impact of deleterious variation in livestock. PhD thesis, Wageningen University, the Netherlands

The genome provides a blueprint of life containing the instruction, together with the environment, that determine the phenotype. In animal breeding, we try to understand this relationship between an animals genetic code (genome sequence) and its performance (phenotype) to select the best performing animals for the next generation. Recently, rapid improvements in genome sequencing have opened up new possibilities to explore and use the complete set of genetic variation seen in (livestock) animals that can be exploited by scientists to try and further close the genotype-phenotype link. However, despite the vast increase of molecular data, pinpointing the exact variants underlying a phenotype of interest is still challenging. In this thesis I provide an in-depth analysis of population genomics and transcriptomics data to identify deleterious and functional variation in livestock populations. More specifically, I report several variants that causes lethality in homozygous state within different stages of development. Moreover, I pinpoint the exact causal mutations and describe its functional consequences at the molecular, phenotypic, and population level. Subsequently, I focus on the identification of functional variation underlying important selection traits. I combine various sources of functional (epi)genomic data to predict the impact of variation in livestock. Together I provide a comprehensive overview of high-impact variation and molecular mechanisms affecting important phenotypes in various livestock breeds, and discuss that molecular genomics could benefit genomic prediction in livestock.

Contents

1

General introduction

1.1 Genome variation

The central dogma in molecular biology describes a process by which DNA is transcribed to RNA which is translated to proteins (Crick 1970). The proteins perform most of the work in our cells, which in return are responsible for the regulation, structure, and functioning of the individuals tissues and organs. The DNA provides a blueprint of life containing the instruction to grow, develop, survive, and reproduce. The complete genetic code (i.e. the genome sequence) defines the phenotype of an individual (Lehner 2013). Hence, the differences (i.e. genetic variation) in genome sequences contribute to the observed phenotypic variation in a population.

Animal breeders try to utilize variation in genomes to select the best performing animals for the next generation (Georges et al. 2019). However, until recently, the characterization of the entire genome sequence of an individual was a costly and time-consuming procedure (Shendure et al. 2017). Yet, rapid improvements in genome sequencing have opened up new possibilities to explore and use the complete set of genetic variation seen in (livestock) animals. Within the department of Animal Breeding and Genomics at Wageningen University, hundreds of animals from various livestock populations have been sequenced. The sequences can be mapped back to a reference genome, built from a single individual to produce a high quality genome (Hillier et al. 2004; Groenen et al. 2012), to explore the millions of genetic variants present in the population. Most of these variants are substitutions of a single nucleotide that occurs at a specific position in the genome (SNPs), or short deletions or insertions (indels). Until recently, the livestock genomes were of much lower quality compared to the human or mouse genome. With the developments in long read sequencing technology, much improved reference genomes for chicken, cattle, and pig have been created (Warren et al. 2017; Warr et al. 2019). These improved reference genomes have significantly improved variant discovery and further downstream analysis.

1.2 From sequence to phenotype

Domestication constitutes a long-term biological experiment, started by early farming societies and currently done by animal breeders, to constantly improve stock. The ultimate goal in livestock genomics is to understand the relationship between the genome of the animal and its phenotype (i.e. the performance). However, this link between an animals genetic code, and the performance is extremely complex (Gjuvsland et al. 2013).

A major breakthrough in animal breeding was provided with the discovery of genomic selection, at the start of this century (Meuwissen et al. 2001), which led to enhanced rates of genetic improvement and shortening of generation intervals. In general, genomic selection uses a variant panel on a chip (SNP chip) that is distributed across the genome, allowing to capture within-breed genetic variation used to select the best animals for breeding. By that, a wealth of genomic information has become available for many species, including livestock (Dekkers 2012). However, genomic selection uses the genome as a black box, as the SNPs on the chip are not causal, but genetically linked to the actual causal variants and genes (Habier et al. 2013).

Despite the remarkable success of genomic selection, the molecular drivers underlying the selection traits are still largely unknown. One factor that explains the success of genomic selection is that most important traits (e.g. growth, fertility) are actually regulated by a large set of genes (Hayes and Goddard 2001), with often subtle effects (polygenic traits), while only a few important traits are inherited by a variant in a single gene, or a small subset of genes with large effects (monogenic traits). To unravel the genetic factors underlying these mono- and polygenic traits requires both genomic and phenotypic information to associate regions in the DNA with important traits, the so-called quantitative trait loci (QTL). Although thousands of such QTL have been reported by now (Hu et al. 2019), identification of the causal genes and variants underlying these QTL is still a major challenge. Hence, only few functional variants have been discovered that are currently used in animal breeding (Goddard et al. 2016).

An additional type of functional (high-impact) variants are those that exhibit harmful effects (i.e. deleterious). Inferring the function of deleterious variants (e.g. underlying recessive defects), and underpinning the genetic basis of the disorder is still a major challenge. However, variants underlying severe phenotypes can likely be identified more easily compared to variation with small effects on important selection traits, thereby providing a starting point to study the genotype-phenotype link in livestock.

Glossary

SNP: Single nucleotide polymorphism, resulting in another nucleotide at one single base in the genome.

Indel: Small deletion or insertion of bases in the genome.

Haplotype: A group of alleles in an organism that are inherited together from a single parent.

Genetic drift: The random fluctuations of allele frequencies from generation to generation.

Mutation load: The total genetic burden in a population resulting from accumulated deleterious mutations

Recessive lethal: A variant resulting in a lethal phenotype in homozygous state (two copies of the variant)

SNP array: A type of DNA microarray which is used to detect polymorphisms within a population

WGS: Whole genome sequence

Effective population size: The number of individuals in a population who contribute offspring to the next generation

LoF variant: A variant that lead to a loss of function of a gene and the associated protein.

Missense variant: A point mutation in which a single nucleotide change results in a codon that codes for a different amino acid

1.3 Deleterious alleles

In animal breeding and genomics we try to predict the effect of specific genetic variants on the phenotype or performance of farm animals. The genetic variants can usually be classified in three categories (though context-dependent): (1) Variants with positive effects; (2) variants that are benign (neutral); and (3) variants with deleterious effects. With the emergence of noticeable genetic defects and a decline in fertility in various important livestock breeds (Cole et al. 2018), inbreeding (and deleterious variants) have now gained large interest among breeders. However, the level of genomic deleterious variation present in a particular breed depends on the past and present population structure and demography. In livestock species, the level of genomic variation is considerably higher compared to human, suggesting a large reservoir of both positive, benign, and deleterious variation in livestock populations (Groenen et al. 2012; Charlier et al. 2016). However, most deleterious alleles generally occur at low frequency, and their harmful effects are only expressed in homozygotes, usually only exposed by inbreeding, caused by inherited copies of genetic material from a recent common ancestor (Bosse et al. 2018). Hence, the set or rare variants in a population are often enriched for deleterious variants, while higher frequency variants are likely enriched for benign or positive variants. Moreover, most deleterious alleles only have mildly deleterious effects, i.e. only contributing to the phenotypic variation as such (Altshuler et al. 2012). However, some variants do have major impact, severely affecting individual fitness (Georges et al. 2019). These major impact variants are usually recessive, and the effect is only exposed in homozygous state. The majority of the currently known deleterious variants characterized in livestock populations directly affect the protein sequence. These can be readily identified using generic annotations of whole genome sequence data. However, other deleterious alleles, affecting the regulatory regions of the genome, are less well characterized, and require novel tools to be properly characterized (Zappala and Montgomery 2016).

The number and frequency of deleterious genetic variants in a population is affected by several factors, genetic drift (i.e. the random fluctuation of allele frequencies in the population), mutation rate, and selection (Charlesworth 2009). The use of a small number of elite breeding-sires has caused a reduction in effective population size (Ne) in many livestock populations (i.e. the number of individuals in a population who contribute offspring to the next generation) (Hall 2016). As a result, small effective population size and artificial selection can impact population fitness substantially, and can lead to a higher risk of inbreeding depression (Howard et al. 2017b). Inbreeding depression is the result of the accumulation of deleterious alleles that increase in frequency (Charlesworth and Willis 2009), mainly due to genetic drift. Other factors also contribute to the (local) landscape of deleterious alleles in a population, including recombination and genetic hitchhiking, which is a change in allele frequency due to the allele being passed along together with a variant that is under selection (Chun and Fay 2011). Therefore, the role of random drift and selection in increasing the frequency of deleterious variants is complex. But certainly, when effective population size is small, drift effects can result in less effective selection (Jimenez-Mena et al. 2016), and some deleterious alleles will drift to relatively high frequency just by chance.

1.3.1 Domestication has influenced the landscape of deleterious alleles in livestock

Domestication of farm animals has coincided with severe population bottlenecks that played a major role in shaping the landscape of deleterious alleles in livestock populations. The domestication bottlenecks have resulted in elevated proportions of deleterious genetic variation in domestic animals, compared to their wild counterpart (Makino et al. 2018). An exception is the European domestic pig, which show no signs of a strong population bottleneck, likely due to a continuous gene flow from wild populations. Moreover, European pig populations underwent continues human-mediated introgression of Asian haplotypes over the last centuries (Bosse et al. 2014), which means most commercial pig populations are essentially hybrids between two subspecies. Hence, despite the bottlenecks associated with domestication, the general population genomic diversity and fitness does not need to decline as a result of domestication and selection.

1.3.2 Detection of deleterious alleles from sequence data

With the emergence of next generation sequencing technologies, the functional impact of alleles, in particular deleteriousness can potentially be assessed. However, the identification of deleterious alleles from WGS data is currently mainly focussed on variants that alter the protein, affecting the coding sequence of the gene. Tools like Sift (Kumar et al. 2009) and PROVEAN (Choi and Chan 2015) are regularly used to score the deleteriousness of missense variants, caused by a single nucleotide change resulting in a codon that codes for a different amino acid in the protein. A second class of deleterious variants are those that cause a loss-of-function (LoF) of the protein. This class includes variants affecting RNA splicing on the splice-donor or splice-acceptor dinucleotide sites, prone to disrupt proper splicing of the mRNA (Lewandowska 2013). Secondly, small deletions or insertions affecting the coding sequence can lead to a frameshift, resulting in a truncated or malformed protein (Watson 2014). Also variants affecting the start codon, or variants that induce a premature stop codon (stop-gained variant) usually cause a loss-of-function of the protein (Watson 2014). The missense and LoF variants can only be annotated in the coding region of the genome, which only accounts for about 3% of the whole genome.

Variants outside of the coding regions can also have a high impact. It is expected that the majority of phenotypic changes due to selection are due to changes in gene expression, rather than protein changes. Gene expression is mostly regulated by transcription factors and local epigenetic modifications in the DNA. Deleterious variants affecting (regulatory) sequences in the non-coding part of the genome, altering gene expression, have remained, so far, undetected. To solve this coding-sequence bias for deleterious alleles, in human genetics, several tools have been developed to score the deleteriousness of any possible variant in the genome, including those in the non-coding part. One such popular tool is Combined Annotation-Dependent Depletion (CADD) (Rentzsch et al. 2019), providing impact scores of any nucleotide substitution in the genome. The CADD framework is built on many features of annotations including the sequence context, conservation scores, gene expression data, protein scores, and epigenomic data. One of the key features utilized by the CADD approach are the sequence conservation scores. Basically, the higher the evolutionary conservation of a site, the likely more important it is for the fitness of the organism. Hence, these conserved sites are sensitive to change, while sites that are less conserved are likely tolerant to change. The CADD scores can prioritize functional, deleterious and disease causing variants over the entire genome in human genetics.

1.4 Purifying selection

Deleterious alleles are generally held at low frequency by a balance between the rate at which they arise by mutation, and the effectiveness of purifying selection at removing them from the population (mutation - selection balance) (Hedrick and Garcia-Dorado 2016). The efficiency of purifying selection depends on the fitness effects of the deleterious alleles, but also on the effective population size, N_e, affecting the influence of genetic drift. For example, a severe and sudden population decline increases drift effects, boosting the frequency of some deleterious alleles in the population. However, for populations that undergo a more gradual population decline, deleterious alleles can be removed more efficiently (Hedrick and Garcia-Dorado 2016). Most commercial populations are under long term selection, with a stable but generally low N_e, suggesting that most harmful deleterious alleles could be efficiently purged. However, since most deleterious alleles are of low impact (and will therefore have low selection coefficient), the majority of deleterious alleles will be very difficult to purge. In addition, strong artificial selection could also lead to strong hitchhiking effects, i.e. deleterious alleles hitch hike along with alleles that are heavily selected, causing rapid increase in frequency (Barton 2000). Also recombination plays an important role to purge deleterious alleles, which is more effective in regions with high-recombination compared to regions showing low recombination levels (Stapley et al. 2017).

1.4.1 Inbreeding

Although deleterious alleles are generally widespread throughout populations, their effect is usually low because the overwhelming majority of these alleles are rare in a population. However, within domestic and wild populations that have small N_e, those alleles might be exposed due to inbreeding, caused by matings between related parents that inherited the same recessive deleterious allele from a recent common ancestor (Howard et al. 2017b). High levels of inbreeding, together with low recombination rate could lead to long stretches of homozygosity in the genome (i.e. runs of homozygosity) (Bosse et al. 2012), enriched for homozygous deleterious variants (Szpiech et al. 2013). In general, runs of homozygosity tend to become smaller over time (if further inbreeding is prevented), as a result of recombination and purifying selection, while recent inbreeding coincides with long runs of homozygosity (Curik et al. 2014). Inbreeding can ultimately result in a general fitness decline of the population (inbreeding depression). Hence, animal breeders try to minimize inbreeding by looking at kinship between selection candidates, but conceding as little as possible on the genetic gain (optimal contribution, Leroy 2014). Allowing for more inbreeding leads to stronger selection response on the short term. However, increased inbreeding coincides with a general loss of genetic diversity, which likely leads to decreased selection response on a longer term (Mulder et al. 2019). Hence, animal breeders generally try to find a balance between genetic gain and the preservation of the genetic diversity within a breed.

1.4.2 Balancing selection for deleterious alleles

The number of deleterious alleles arising from mutation is expected to be equal to the number of deleterious alleles purged by (natural) selection, resulting in a mutation - selection balance (Charlesworth and Willis 2009). However, this equilibrium does not seem to apply to most livestock populations. First of all, the effect of genetic drift is relatively large, because of low N_e. Secondly, more evidence arises that several (deleterious) alleles can be under balancing selection, for example, when heterozygotes exhibit higher fitness compared to homozygotes. In cattle, various examples of balancing selection have been described, driving deleterious alleles to higher population frequencies (Kadri et al. 2014). This effect of balancing selection is likely further driven by crossbreeding, since in the (terminal) crosses the alleles only appear in heterozygous state, likely underlying heterosis to some degree (Charlesworth and Willis 2009). Also, in commercial breeding, the breeding goal is often determined by a combination of several traits, weighted on its importance. Many of these traits are negatively correlated, for which alleles could have pleiotropic effects on different traits in the population, positively affecting one trait, while negatively affecting another trait (Xiang et al. 2017).

1.5 Recessive lethals

1.5.1 Impact of selection and drift

Recessive lethals cause mortality in homozygous individuals by affecting genes important for growth or development, which mostly manifest early in development (Gluecksohn-Waelsch 1963). Although lethal alleles are present in every population, the population structure and N_e have major impact on the frequency of lethal alleles; Small N_e leads to high extinction rates of de novo recessive lethals, but the few that are not lost tend to spread and increase in frequency (Charlier et al. 2016). The impact of genetic drift on recessive lethals is slightly different compared to mildly deleterious or neutral alleles, because of the absence of homozygous individuals in a mature population. Therefore, at a certain frequency, an equilibrium between drift and selection is reached (i.e. the maximum allele frequency solely reached by drift), at which the loss of homozygotes will prevent further increase assuming no heterozygote advantage.

Most lethal recessives are thought to cause early lethality (i.e. in embryonic state), knocking out essential genes required for early embryonic development. However, other lethals might be expressed later in gestation, or postnatally, usually resulting in bigger economic loss for the breeders. In general, lethal recessives go unnoticed in the population, caused by the low-frequency nature, and marginal effect on fertility, exhibited only in carrier-by-carrier matings. Even recessive defects that reveal themselves postnatally (i.e. causing juvenile mortality) often go unnoticed because of unawareness of genetic defects. Furthermore, many rare and indistinct defects often are unknown to have a genetic basis. Fortunately, the effects of recessive lethals are usually masked by crossbreeding, unless the same allele is segregating in multiple breeds, used for generating crossbred animals (Cassady et al. 2002). Nevertheless, to effectively select against specific low-frequency syndromes in the purebred populations does require new approaches.

1.5.2 Methods to identify lethal recessives

Genomic selection is not very efficient in eliminating rare deleterious variation because the low-frequency alleles are not well captured by the SNPs on the selection chips (Zhang et al. 2019). Moreover, the recessive deleterious effects are only exposed in carrier-by-carrier (CxC) crosses, which only seldomly occur for low-frequency alleles. However, the availability of a large number of genotyped, pedigreed individuals enables the unravelling of the genetic basis of rare disorders in the population by applying alternative methods. One of such method is to test for statistical depletion, or even the absence, of specific haplotypes in homozygous state in the population. This haplotype approach is a tool originally developed for cattle by vanRaden et al. (2011). The power of this method heavily depends on the number of genotyped individuals. If applied to tens of thousands of genotyped animals, a number now routinely attained in many commercial breeding lines, even very rare deleterious haplotypes (frequency < 2%) can be detected. An alternative method that does allow such rare deleterious alleles to be identified is to sequence the entire genome of tens to hundreds of animals from a single population (Charlier et al. 2016), and to identify potential phenotype-altering variants from the sequence, ranging from embryonic lethal to only mildly deleterious mutations. Hence, population sequence data is potentially more powerful to detect deleterious alleles, allowing to identify functional genomic information that can be utilized in breeding. However, to assign function to (sequence) variation remains challenging, and requires novel bioinformatics approaches to tackle (Willet and Wade 2014).

1.5.3 Essential genes

The impact of a genic variant also depends on the gene that is affected. Not every gene is essential, and some large gene families are prone to be affected by loss-of-function variants, leading to pseudo-genes, that lost its functionality. Studies in other mammals can be used to infer which genes are essential, in particular natural knock-outs in humans, and experimental knock-out systems (especially in rodents). In human, by sequencing thousands of human exomes, genes have been scored for their tolerance to loss-of-function mutations (Petrovski et al. 2013). Basically the observed rate of LoF variants given all the protein-coding variation found in the gene indicates whether a gene is tolerant to LoF variants, marking whether it is essential or not. Genes that are least tolerant to LoF mutations are the genes for which a single copy is not sufficient to produce a normal phenotype (i.e. leading to haploinsufficiency), while other genes need at least one working copy, that could be affected by recessive deleterious variants in homozygous state. A third class of genes are completely tolerant to LoF mutations, not leading to a notably reduced fitness if knocked-out. In mice, a large number of experimental gene-knockouts have been performed. The results are summarized in the Mouse Genome Informatics (MGI) database, providing a comprehensive resource for mice knockout phenotypes (Blake et al. 2017). The information from these human and mouse studies can be applied in livestock genome studies, since mammals share many conserved basic developmental pathways (Basson 2012).

1.6 Functional variation

1.6.1 Identification of selective sweeps

Animal breeding has resulted in strong artificial selection on important traits in the breeding program (Andersson 2012). During selection, a number of beneficial variants with high impact have been selected to become fixed (i.e. homozygous) in the population. These regions where these high-impact variants reside show the result of ‘hitch-hiking’ and are usually referred to as “selective sweeps”. The hitchhiking effect results in a reduction of genetic variation near the beneficial locus, as a consequence of strong selection on it (Andersson 2012). Whole genome sequence data has led to the discovery of variants that are beneficial in commercial livestock population, e.g. by looking for regions in the genome that are under strong (recent) positive selection (Rubin et al. 2010; Rubin et al. 2012). A number of strong sweeps have been identified affecting genes responsible for important traits in livestock. However, an ongoing challenge for this approach is to differentiate true selected variants and variants that increased in frequency as a result of genetic drift.

1.6.2 Phenotypic variation explained by variants affecting regulatory sequences.

The variants with large effects underlying strong selection signatures represent only a fraction of the total functional variants in the genome. This is because most important selection traits are regulated by a large number of genomic loci, any one with relative minor effects, not leaving clear selection signatures in the genome. Moreover, it is now completely evident that changes in gene expression, caused by changes in regulatory sequences, have a much larger impact on selection traits, whereas changes in the coding sequence only contribute sporadically (Georges et al. 2019).

To further close the genotype-phenotype gap in livestock, genome wide association studies (GWAS) are now routinely performed to associate genomic markers to important traits in livestock (Sharma et al. 2015). Despite the relative success of identifying QTL using GWAS approaches, the identification of the actual causal variants has been far less successful. Partly because the resolution of GWAS is limited by the correlation between neighbouring markers in linkage disequilibrium (LD), and because the causal variants often reside in the noncoding regions of the genome, enriched in predicted transcriptional regulatory regions affecting gene expression (Ponting and Hardison 2011). Therefore, further functional genomic and transcriptomic information will be key to pinpoint causal variation in livestock (Ron and Weller 2007).

However, in livestock, the level of functional genomics information is rather limited. To improve this, the FAANG (Functional Annotation of Animal Genomics) consortium is running several projects to characterize the functional genome in various animal species, focussing on livestock (Giuffra et al. 2019). In addition, applying techniques such as CADD in livestock will provide much improved functional predictions of the variation, especially in the non-coding part of the genome. Hence, the data and framework that will be generated within the FAANG consortium will accelerate the discovery of functional variants in livestock species, and will provide the basis for generating livestock specific CADD tools. Applying selection directly to the causative variants will enhance the efficiency of selection (Xiang et al. 2019), establishing the use of functional genetic variation in genomic selection.

1.7 Thesis outline

The main goal of my research is to further close the genotype phenotype gap in livestock. First I investigate deleterious alleles and recessive lethal alleles in purebred livestock breeds. Next, I investigate the functional basis for the deleterious variants, and the forces that maintain the deleterious variants in the population including drift and balancing selection. Finally, with a comprehensive set of functional annotations, I attempt to go beyond deleterious, and predict functional variation underlying important phenotypes in commercial livestock breeds. In **chapter 2** I perform a survey to assess deleterious haplotypes that likely harbour a recessive lethal allele in three pig populations. **Chapter 3** describes the causal mutation for feathering rate in turkey, which affects the same gene responsible for the slow-feathering phenotype in chicken, but with a different mutation. **Chapter 4** provides a genomic perspective on deleterious and functional genetic variation in three egg-laying breeds, giving insight into the process of purifying selection for breeds under strong artificial selection. In **chapter 5** I describe a large deletion under balanced selection that affects the *BBS9* gene inducing late fetal mortality in homozygous animals, while having positive effects for growth and feed intake in heterozygotes. **Chapter 6** describes loss of function mutations in essential genes that cause embryonic lethality in pigs. I reflect on the effect of genetic drift on lethal recessive variants, and discuss the impact of the lethals on population fitness, and its role in the heterosis effect observed for crossbred litters. In **chapter 7** I describe a recessive 16-bp deletion in the *SPTBN4* gene causing severe myopathy and postnatal mortality in pigs. **Chapter 8** provides a framework (using pig CADD scores) to pinpoint likely causal variation and genes underlying important phenotypes in pigs. Finally, the applications and implementations of the findings are discussed in **chapter 9**, and several guidelines are discussed to prevent the consequences of such harmful mutations.

2

A systematic survey to identify lethal recessive variation in highly managed pig populations

Martijn F. L. Derks[1], Hendrik-Jan Megens[1], Mirte Bosse[1], Marcos S. Lopes[2,3], Barbara Harlizius[2], Martien A. M. Groenen[1]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. [2]Topigs Norsvin Research Center, Beuningen, the Netherlands. [3]Topigs Norsvin, Curitiba, Brazil

BMC Genomics (2017) 18:858

Abstract

Lethal recessive variation can cause prenatal death of homozygous offspring. Although usually present at low-frequency in populations, the impact on individual fitness can be substantial. Until recently, the presence of recessive embryonic lethal variation could only be measured indirectly through reduced fertility. In this study, we estimate the presence of genetic loci associated with both early and late termination of development during gestation in pigs from the wealth of genome data routinely generated by a commercial breeding company.

We examined three commercial pig (*Sus scrofa)* populations for potentially deleterious genetic variation based on 80 K SNP-chip genotypes, and estimate the effects on reproductive traits. 24,000 pigs from three populations were analyzed for missing or depletion of homozygous haplotypes. We identified 145 haplotypes (ranging from 0.5–4 Mb in size) in the genome with complete absence or depletion of homozygous animals. Thirty-five haplotypes show a negative effect on at least one of the analysed reproductive traits (total number born, number of stillborn, and number of mummified piglets). One variant in particular appeared to result in relative late termination of development of fetuses, responsible for a significant fraction of observed stillborn piglets ('mummies'), as they die mid-gestation. Moreover, we identified the *BMPER* gene as a likely candidate underlying this phenomenon.

Our study shows that although lethal recessive variation is present, the frequency of these alleles is invariably low in these highly managed populations. Nevertheless, due to cumulative effects of deleterious variants, large numbers of affected offspring are produced. Furthermore, our study demonstrates the use of a large-scale commercial genetic experiment to systematically screen for 'natural knockouts' that can increase understanding of gene function.

Key words: Population genomics, Genetics, Deleterious variation, Embryonic lethality, Mummified piglets

2.1 Introduction

Small effective population size can lead to inbreeding depression. The cause of inbreeding depression is the accumulation of (recessive) deleterious alleles increasing in frequency and becoming expressed in homozygous state due to drift in small populations (Charlesworth and Willis 2009). In domesticated populations, despite strong artificial selection for desired traits, selection on relatively rare variation (allele frequency < 10%) is usually very inefficient (Kearney et al. 2009). The eradication of deleterious variation is challenging by applying either traditional or even more recent ‘genomic’ breeding strategies. The inefficiency of purging deleterious variation, even from highly managed populations, is particularly apparent if there is an unpredictable, or poorly characterized relationship between genotype and phenotype. For instance, when a homozygous deleterious phenotype leads to very early death of the developing embryo, the only observed consequence is a (somewhat) lower fertility of the parents. It is estimated that livestock species harbour 2–4 fold higher variation in the genome compared to humans (Bosse et al. 2014; Charlier et al. 2016). Accordingly, the number of non-synonymous mutations has been shown to be greater in livestock animals, suggesting a large reservoir of potentially deleterious variation in livestock (Charlier et al. 2016).

Domestication in general, and modern breeding industry in particular, constitutes the largest and longest lasting genetic experiment ever conducted. In many cases, especially in pig and poultry breeding, commercial breeding organizations apply their genetic improvement efforts at a small number of elite breeding lines (Gonzalez-Pena et al. 2015). These breeding lines are usually fairly closed, i.e. exchange between breeding lines is infrequent. Moreover, depending on species and breeding purpose, the effective populations sizes vary from small (several hundred at most) to very small (dozens of animals). With the adoption of genomic selection, a large proportion of the animals in the pure bred elite lines, i.e. the selection candidates, are genotyped using high to medium density SNP assays to estimate their genomic estimated breeding value (GEBV). Breeding values itself are not very efficient in eliminating rare deleterious variation. However, the availability of a large number of genotyped, pedigreed individuals enables the unravelling of the genetic basis of rare disorders in the population. Recessive deleterious variants can be identified by testing for statistical depletion, or even the absence, of specific haplotypes in homozygous state. This haplotype approach is a powerful tool (VanRaden et al. 2011; Fritz et al. 2013; Sahana et al. 2013; Pausch et al. 2015; Haggman and Uimari 2016) originally developed for cattle by vanRaden et al. (2011). The power of this method heavily depends on the number of genotyped individuals. If applied to tens of thousands of genotyped animals, a number prohibitively large usually for academic budget, but routinely attained in many commercial breeding lines nowadays, even very rare deleterious haplotypes can be detected (frequency < 2%). Genotyping large numbers of domestic animals currently still relies on the use of dedicated SNP assays ('SNP chips'). Since these assays are designed with a bias towards high minor allele frequency (MAF), causal variants of serious syndromes are unlikely to be present on the assay. The haplotype based approach is therefore more efficient in capturing deleterious variation compared to individual SNPs, which are expected to be often in low linkage disequilibrium (LD) with the causal variant. Significant depletion of haplotype homozygosity is an indication of decreased viability, and these haplotypes are likely to harbour deleterious mutations causing embryonic lethality (EL) in homozygous state. Studies in mouse showed a lethal knockout phenotype for about 30% of mouse genes in homozygous state, suggesting a large proportion of potential embryonic lethal genes in Mammalia (Ayadi et al. 2012).

Fertility traits, such as total number born (TNB), are among the relevant phenotypes systematically recorded in pig breeding that can provide phenotypic support for (early) embryonic lethality. In addition, recorded phenotypes, such as number of stillborn (NSB) and the number of mummified pigs (MUM), can provide additional information on genetic defects that are lethal, or seriously compromising the survival probability later in foetal development. In theory, assuming prenatal death of homozygotes, a loss of 25% of each litter is expected if two heterozygous carriers of the lethal variant mate (C x C mating). In mammals, death of an embryo or foetus usually does not result in spontaneous termination of the pregnancy, when also carrying living young. Instead, foetuses go through a process of desiccation and encapsulation, known as mummification (Christianson 1992). Foetal mummification can occur from day 35 of gestation until parturition (when the skeletal system is developing), and has been associated with several risk factors like large litter size and infectious disease (Dron et al. 2014). Also, several studies have already identified QTLs associated with pig reproductive traits, including mummified piglets (Onteru et al. 2012; Schneider et al. 2012; Hernandez et al. 2014; Verardo et al. 2016). Parameters, such as time of death and any morphological abnormalities found in mummified piglets (Christianson 1992), can provide insights in the developmental consequences of a specific genetic defect without any further welfare concerns for the mother or live siblings. A systematic analysis to identify specific genetic defects as risk factors, however, has not been conducted.

In this study, we aim to identify novel genetic loci associated with both early and late termination of development during gestation in pigs. Moreover, we examine the occurrence and frequency of lethal haplotypes in the studied breeds, as well as their impact on fertility related traits. Finally, we show that missing homozygosity in highly managed livestock populations can be a result of early lethality caused by low frequency recessive lethal haplotypes.

2.2 Results

2.2.1 Screening for haplotypes exhibiting missing or deficit homozygosity

Breeding lines are used in three- or four way crosses to produce large numbers of slaughter pigs (Gonzalez-Pena et al. 2015). However, elite breeding lines are generally kept as closed populations, and selection is done within these populations. Because of this characteristic, these breeding lines meet two criteria for the method applied in this study to be successful: a) we can expect that not all deleterious variation is effectively purged from the population, and that low to moderate allele frequencies for some deleterious variation remains in the population, and b) because we specifically examine C x C matings, we expect 25% of the offspring to be homozygous for the carrier haplotype, a necessary prerequisite when scanning for missing homozygotes. In total, we scanned for missing homozygosity in 5517 pigs from a synthetic elite sire (BR) line with Large White and Piétrain genetic background, 5301 Landrace (LR) and 12,982 Large White (LW) pigs, the latter two representing two elite dam lines. The dam and sire lines are selected towards distinct breeding goals. Dam populations are primarily selected for female reproductive traits, whereas sire lines are primarily selected for production traits. We found no evidence of (recent) inbreeding in any of the three lines (F-coefficient close to zero, Additional file 1: Figure S1). The statistical power of our study stems from a total of 23,800 animals from the three pure lines, genotyped on low to medium (10 K, 60 K, and 80 K) density SNP arrays (Additional file 1: Table S1-S2). Animals genotyped on either the 10 K or 60 K panel were imputed to 80 K with generally high accuracies (Additional file 1: Table S3). After filtering, a final set of 22,961 animals was used for further analysis (Additional file 1: Table S4). After haplotype phasing, we systematically examined the genome by using an overlapping sliding window approach to assess haplotype frequencies. Haplotypes were marked as potentially deleterious if a significant deficit (exact binomial test) in homozygotes was observed.

We identified 22, 10, and 56 haplotypes with missing homozygosity (MH), and 19, 6, 32 haplotypes exhibiting a statistically significant deficit homozygosity (DH) in the BR, LR, and LW line, respectively (Table 2.1, Additional file 1: Figures S2-S4**,** Additional file 2). DH haplotypes have either incomplete LD with the causal variant or incomplete penetrance of the variant at the phenotypic level in homozygous state. The haplotype lengths varies from 0.5 to 4 Mb and frequencies range from 0.8 to 11.4% for haplotypes with MH and from 2.6 to 15.8% for haplotypes with DH. The larger number of genotyped animals and trios (both parents and offspring genotyped) in the LW breed allowed for the identification of a high number of low frequency haplotypes (< 3%), compared to the other two breeds (Additional file 1: Figure S5). For the haplotypes with significant depleted homozygosity, the number of expected homozygotes ranged from 5.75 to 140.25 animals per haplotype, with an overall average of 18.4, 22.2, and 24.3 expected homozygotes for LR, LW, and BR breeds, respectively (Additional file 1: Figure S6). We expect a larger proportion of heterozygous carriers from C x C litters due to the missing homozygote offspring. Hence, the percentage of heterozygous carrier offspring is greater than 50% for the majority of the haplotypes in all three breeds (Table 2.1 **,** Additional file 2).

Table 2.1: Description of the data for missing and depleted homozygous haplotypes in three pig breeds. Table shows average and standard deviation (between parenthesis) for all parameters per breed. The number of loci harbours the unique number of genomic windows containing significant haplotypes.

Description	**Synthetic boar line**	**Landrace**	**Large white**
Number of samples	5,488	5,056	12,417
Number of trios	3,806	2,548	8,778
Number of haplotypes	41	16	88
Number of loci	32	16	70
Haplotype length (markers)	25.4 (18.5)	19.3 (22.3)	36.04 (37.0)
Haplotypes in window (frequency > 0.5%)	16.3 (7.8)	22.2 (10.3)	25.1 (13.3)
Number of carriers	707.5 (348.9)	689.9 (290.8)	972.1 (545.3)
Haplotype frequency	6.4 (3.2)	6.8 (2.9)	3.9 (2.2)
Homozygous expected	24.3 (27.6)	18.4 (10.5)	22.2 (22.6)
Carrier matings with genotyped offspring	29.9 (31.6)	36.1 (20.6)	33.6 (34.6)
Carrier matings in pedigree	72.9 (69.5)	169.8 (119.0)	104.7 (107.8)
Genotyped carrier progeny	96.8 (109.7)	73.75 (42.1)	88.8 (90.2)
% Heterozygous carrier progeny	55.1	60.3	53.8
Genes in window	21.2 (26.3)	15.8 (22.0)	19.0 (19.7)

2.2.2 Genomic regions enriched for missing or deficit homozygosity

We identified four regions enriched for MH and DH haplotypes over all three breeds: SSC1:294–297.5, SSC2:156–159.75, SSC3:142–144, and SSC11:70–73.5 Mb (Additional file 3). Together these four regions account for 42 of the total of 145 identified haplotypes identified in all three lines. These loci have previously been identified as copy number variable regions (Paudel et al. 2015), but were not extensively linked to reproductive traits (Hu et al. 2016). A gene-set enrichment analysis for the genes overlapping the 145 identified haplotypes revealed only one significant annotation cluster, i.e. related to olfactory receptors (DAVID enrichment score 3.45) (Additional file 3, Table S5).

2.2.3 Association with reproductive traits and candidate gene selection

We examined all 145 significant haplotypes for their effect on three reproductive traits: TNB, NSB, and MUM. Phenotypic records for all three traits were available for both dam lines. For the boar line, only records on TNB were available. We listed all phenotypes from C x C matings and carrier x non-carrier matings (C x NC) to identify missing or depleted haplotypes affecting these traits. Haplotypes significantly affecting fertility are named and ranked according to breed, affected phenotype, and genomic location. Figure 2.1 shows the genomic distribution of the haplotypes affecting fertility per breed.

2.2.4 Total number born

We identified 26 haplotypes exhibiting a significant reduction in TNB (Table 2.2), three in the BR line, 4 in LR, and 19 in LW. The reduction in TNB ranged from 2.84 to 18.72% representing 0.45 to 2.97 piglets per litter. Candidate genes were identified based on early lethality in knockout mice studies and could be identified for 16 haplotypes (Additional file 4). Fourteen regions were not previously associated with reproductive traits in livestock and can be considered as novel according to the 2016 pig QTL database (Hu et al. 2016) (Additional file 4). Six haplotypes exhibit a large reduction (> 10%) of TNB. LR4, found on SSC13, shows a reduction of 17.13% in TNB based on 44 C x C matings. This 0.5 Mb region contains 12 protein coding genes (Additional file 4), of which *KLHL40* and *POMGNT2* cause early lethality in knockout mice (Eppig et al. 2015). Moreover, haplotype LW9, spanning a 4 Mb region on SSC7, exhibits a reduction in TNB of 15.61% and overlaps with 13 candidate genes that could potentially cause early lethality (Eppig et al. 2015). Haplotype LW14, previously associated with TNB (Schneider et al. 2012), shows a reduction of 11.25% in TNB.

This region contains two candidate genes, one causing early embryonic lethality before time of implantation in mice (*UROS*), and the other causing post-natal lethality (*ADAM12*) (Eppig et al. 2015). Finally, four haplotypes were identified on SSC18, of which LW19 spans a 1 Mb region (43–44 Mb) and exhibits the largest reduction on TNB (18.72%) based on 88 C x C matings. Moreover, this haplotype has been associated with a large increase in the number of mummified piglets and a strong candidate gene could be identified (*BMPER* described below: *LW19 homozygous foetuses become mummified in Large White*).

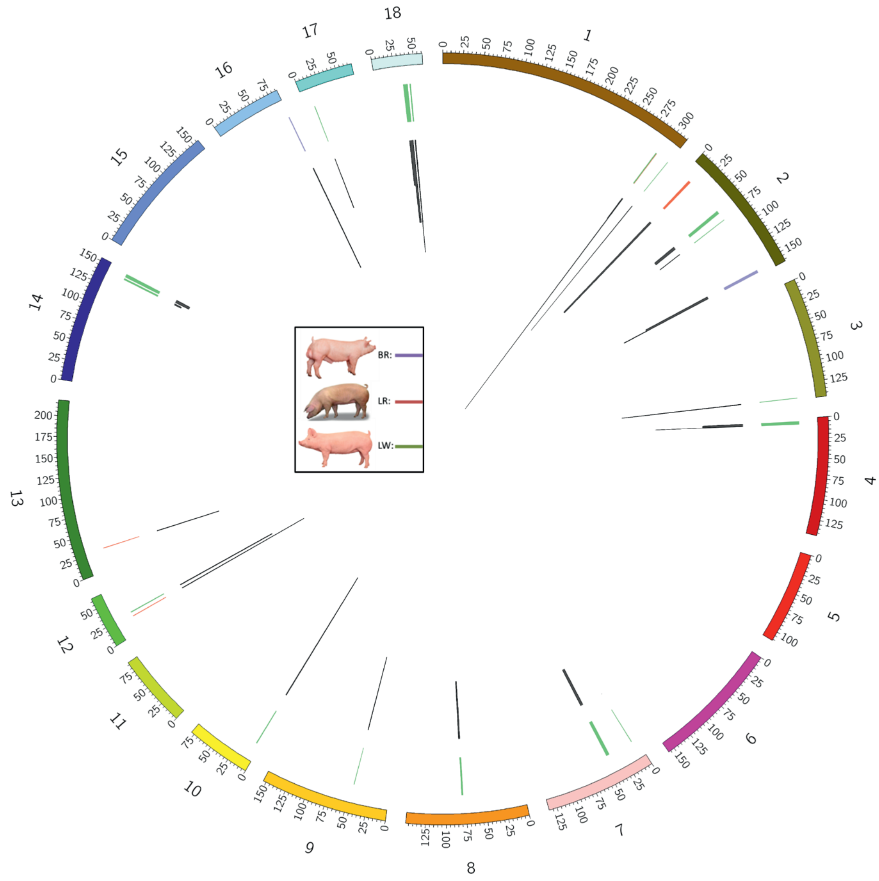


Figure 2.1 Genomic locations of the haplotypes affecting fertility in the BR (purple), LR (red), and LW (green) breed. Figure shows 18 autosomal chromosomes, line width indicates haplotype length. Black lines indicate the relative haplotype frequency ranging from 1.0 to 11.5%. Pig graphics in the figure legend provided by Topigs-Norsvin, all rights reserved.

Table 2.2 Haplotypes affecting TNB. The genomic location and haplotype frequency is provided in columns 1–5. The “homozygotes” section shows expected and observed homozygotes including statistical test. Information on carrier x carrier (C x C) matings and progeny is provided in the “matings” section. Effect on the phenotype is provided in the “reduction in TNB” section.

Abbreviation	Chr	Start	End	Hap. Freq	Homozygotes			Matings			Reduction in TNB	
					Expected	Observed	Exact binomial test	C x C matings	Genotyped progeny	Het. carrier progeny	Percent	P
BR1	SSC2	156	159	3.8	8	0	0.000138	42	32	13	8.192	0.054
BR2	SSC2	158.5	159.5	4.8	18.75	0	5.70E-10	69	75	38	9.220	0.008
BR3	SSC16	85.5	86.5	5.4	17.25	1	8.22E-08	53	69	23	7.037	0.046
LR1	SSC1	295	296	11.4	30.25	0	1.71E-15	370	121	97	3.243	0.012
LR2	SSC2	10.5	13.5	6	6.75	0	0.000585	148	27	18	4.35	0.026
LR3	SSC8	78	79	2.6	8.25	0	0.000144	30	33	17	8.505	0.017
LR4	SSC13	28.75	29.25	3.6	5.75	0	0.002581	44	23	16	17.130	2.16E-06
LW1	SSC1	294.5	295.5	2	7.5	0	0.000138	44	30	20	9.552	0.010
LW2	SSC1	294.75	295.25	11.5	132.75	1	1.87E-72	619	531	418	2.843	0.007
LW3	SSC1	295	295.5	5.9	30.25	0	5.67E-17	259	121	94	4.522	0.003
LW4	SSC1	295	295.5	5.1	37.5	2	1.16E-19	206	150	116	4.213	0.008
LW5	SSC1	313.75	314.25	7.4	126.75	3	3.00E-60	483	507	146	3.875	0.000121
LW6	SSC3	142.75	143.25	5.8	24.25	0	4.27E-14	207	97	31	4.275	0.012
LW7	SSC4	4	8	2.6	26.25	5	7.04E-10	111	105	45	5.311	0.013
LW8	SSC4	6.5	7.5	4.5	14.25	1	4.16E-08	50	57	32	9.854	0.009
LW9	SSC7	40	44	2.6	6.75	0	0.000138	30	27	12	15.612	0.001
LW10	SSC8	78	80	3.3	19.5	5	5.44E-09	173	78	60	4.200	0.023
LW11	SSC9	46.75	47.25	4	6	0	0.00036	56	24	16	7.622	0.012
LW12	SSC10	5.5	6.5	6.5	50.25	5	9.23E-23	296	201	123	4.169	0.003
LW13	SSC12	26.5	27.5	5.2	20	0	9.31E-11	102	80	52	7.383	0.004
LW14	SSC14	146	150	1.6	8	0	0.000144	28	32	15	11.252	0.004
LW15	SSC17	10.5	11	3.1	16	0	3.80E-09	42	64	29	9.920	0.012
LW16	SSC18	34.5	37.5	2.8	6	0	0.000954	30	24	14	11.621	0.027
LW17	SSC18	36	40	4.3	15.25	2	5.22E-07	74	61	28	7.517	0.017
LW18	SSC18	42.75	43.25	5.5	15.75	3	1.32E-07	127	63	43	15.438	1.27E-12
LW19	SSC18	43	44	4.3	11.75	0	1.29E-06	88	47	32	18.715	4.91E-12

Table 2.3 Haplotypes affecting NSB. The genomic location and haplotype frequency is provided in columns 1–5. The "homozygotes" section shows expected and observed homozygotes including statistical test. Information on carrier x carrier (C x C) matings and progeny is provided in the "matings" section. Effect on the phenotype is provided in the "Increase in stillborn" section.

Abr.	Chr	Start	End	Hap. Freq	Homozygotes			Matings			Increase in stillborn	
					Expected	Observed	Exact binomial test	C x C matings	Genotyped progeny	Het. carrier progeny	%	P
LR5	SSC12	21	21.5	6.6	20	0	1.37E-10	189	80	63	29.787	0.016
LW20	SSC2	64	68	2	9.25	0	2.03E-05	160	37	24	32.215	0.007
LW21	SSC2	78.25	78.75	2	8.5	0	5.35E-05	165	34	21	34.228	0.004
LW22	SSC14	142	144	1.3	8.5	0	8.53E-05	20	34	15	57.738	0.038

Table 2.4 Haplotypes affecting MUM. The genomic location and haplotype frequency is provided in columns 1–5. The "homozygotes" section shows expected and observed homozygotes including statistical test. Information on carrier x carrier (C x C) matings and progeny is provided in the "matings" section. Effect on the phenotype is provided in the "Increase in mummified" section. Haplotype already listed in Table 2.2 have similar abbreviations.

Abr.	Chr	Start	End	Hap. Freq.	Homozygotes			Matings			Increase in mummified	
					Expected	Observed	Exact binomial test	C x C matings	Genotyped progeny	Het. carrier progeny	%	P
LW4	SSC1	295	295.5	5.1	37.5	2	3.05E-19	206	150	116	37.209	0.012
LW23	SSC7	6.75	7.25	1	7.25	0	0.000954	17	29	13	256.522	0.028
LW17	SSC18	36	40	4.3	6	2	5.22E-07	30	24	14	224.324	0.044
LW18	SSC18	42.75	43.25	5.5	15.75	3	1.32E-07	127	63	43	375.757	1.67E-10
LW19	SSC18	43	44	4.3	11.75	0	1.29E-06	88	47	32	479.412	2.118E-10

2.2.5 Number of stillborn

Four haplotypes with a significant increase in the number of stillborn were identified (Table 2.3). The increase ranged from 29.8 to 57.7%, representing an increase of 0.42 to 0.97 stillbirths per litter. Candidate genes could be assigned to each haplotype (Additional file 4), but none of the haplotypes has previously been associated with an increased number of stillbirths in livestock according to the 2016 pig QTL database (Hu et al. 2016).

2.2.6 Number of mummified piglets

Analysis of the number of mummified piglets revealed five haplotypes with a significant increase in mummified piglets per litter (Table 2.4). The increase ranged from 37.1 to 479.4%, accounting for an increase of 0.16 to 1.64 mummified piglets per litter for C x C matings compared to C x NC matings. Also, two haplotypes (LW17, LW23) were found in a region previously associated with other reproductive traits (Hu et al. 2016). One of these, LW23, located on SSC7 (6.75–7.25 Mb), shows a 2.5 fold increase in MUM, but no candidate gene could be assigned (Additional file 4). Finally, three haplotypes were identified on SSC18, one of these, LW19, exhibits a complete lack of homozygotes, and shows the largest increase (about 5-fold) in the number of mummified piglets. The two additional haplotypes on SSC18 surrounding LW19 (LW17, LW18), exhibit similar, but less severe, phenotypic effects. We observed a low number of homozygous carriers for LW17 (2 homozygotes) and LW18 (3 homozygotes), suggesting incomplete LD with the causal variant.

Table 2.5: Haplotype LW19 characteristics. Difference is the percentual difference in the average TNB and MUM for C x C and C x NC matings.

Haplotype ID	LW19
Position, Mb	SSC18: 43 - 44
Number of markers	26
Starting marker	ASGA0079708
Ending marker	ALGA0098146
Haplotype frequency %	4.3
Carrier frequency %	8.6
Avg. TNB (difference %)	12.9 (-18.7)
Avg. NBA (difference %)	11.89 (-17.7)
Avg. Mummified (difference %)	1.97 (479.4)
Genes in window	*BMPER, BBS9*

2.2.7 LW19 homozygous foetuses become mummified in large white

Haplotype LW19 (SSC18:43–44 Mb) shows a five-fold increase in the number of mummified piglets, and a 18.71% decrease in TNB calculated from 88 C x C matings (Table 2.5). This locus has not been previously reported to be associated with an increase in the number of mummified piglets. Together these 88 matings produced 173 mummified piglets, 1.97 on average per litter (Table 2.5 , Additional file 1 : Figures S7-S8). Moreover these 173 mummified piglets are responsible for 1.98% of the total number of mummified piglets (8726) recorded for this breed in a decade (December 2006–April 2016). The difference in the ratio MUM/NSB/NBA between C x C (1.97/1.01/11.89) compared to C x NC (0.34/1.43/14.44) is highly significant (P < .0001, Chi-Square). Especially the fraction of litters that contain 2 to 5 mummified piglets per litter is significantly higher for C x C matings (Figure 2.2).

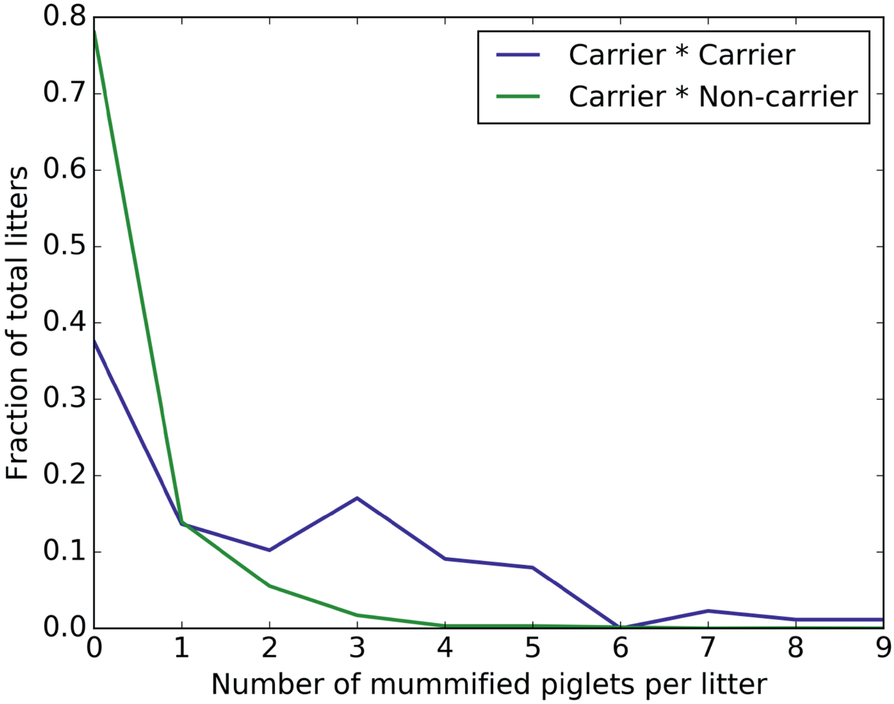


Figure 2.2 Fraction of the number of mummified piglets per litter for haplotype LW19. The axes indicate the fraction of the total litters (y) with a certain number of mummified piglets (x). Figure shows a larger proportion of mummified piglets per litter for the C x C matings compared to C x NC matings, except when n = 1

The carrier frequency for this haplotype is 8.6%, meaning that about 0.74% of the litters in this breed are affected assuming random matings, and 0.185% of all piglets will be affected if penetrance is 100%. We tracked three recent C x C matings with a total of 9 mummified piglets to estimate the approximate age when the foetus has died (an example of a C x C mummified piglet is shown in Additional file 1: Figure S9). The length from crown to rump was about 10–11 cm which corresponds to an age of approximately 55 days (van der Lende and van Rens 2003). The haplotype overlaps with two protein coding genes (*BMPER, BBS9*). *BBS9* has previously been associated with the Bardet-Biedl syndrome in human. This syndrome, however, does not usually cause early lethality (Khan et al. 2016). We therefore focused on the *BMPER* gene as the most likely candidate gene for the observed effect. We performed runs of homozygosity (ROH) and extended haplotype homozygosity (EHH) analysis on SSC18, and identified a region flanking the *BMPER* locus to be potentially under recent positive selection (SSC18:40–43 Mb, Figure 2.3).

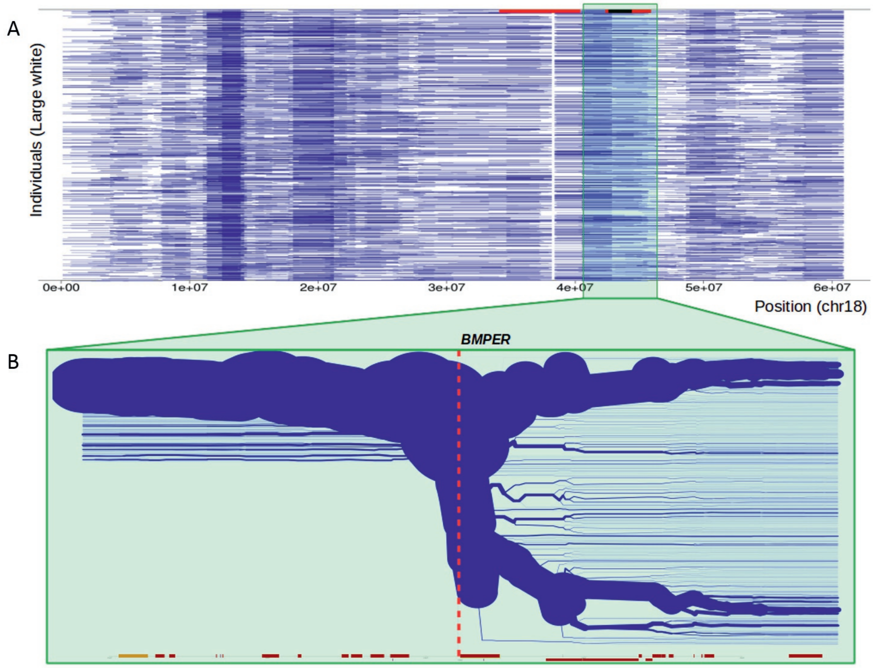


Figure 2.3 Runs of homozygosity (ROH) and extended haplotype homozygosity (EHH) on SSC18. A) Individual Large White pigs are represented as horizontal lines, with blue bars indicating a homozygous segment at that position on SSC18. The red bars on top indicate all significant haplotypes in the Large White, with the haplotype LW19 (SSC18:43-44 Mb) indicated in black. Clustered homozygous segments are an indication of a haplotype putative

under selection. B) Local breakdown of LD in the Large White population at the LW19 haplotype locus. The bifurcation diagram displays haplotypes starting at the *BMPER* locus and extending either up- or downstream of the *BMPER* gene. Line thickness represents proportion of haplotypes. Red and yellow bars indicate locations of genes as annotated in Ensembl (release 87)

2.3 Discussion

Highly managed, domesticated populations are expected to be under selection against inbreeding depression. Indeed, our results show that high frequency occurrence of potentially monogenic lethal or debilitating alleles is rare in commercial populations, despite relatively low effective population sizes. This is in contrast to other domesticated populations that are far less well managed, such as dogs and horses, and that can carry high frequencies of deleterious alleles (Schubert et al. 2014; Marsden et al. 2016). The few examples that exist for commercial breeding populations, mostly from Holstein cattle, indicate that the effects of deleterious traits are often masked, because they involve early embryonic lethality, which reveals itself only indirectly as depressed parent fertility. Moreover, some of these lethal alleles are maintained in the population as a result of balancing selection, where heterozygotes show an advantageous phenotype (Kadri et al. 2014). However, even in those cases, these alleles are usually kept at low frequency.

It is unlikely that purging can remove all or even most of the detrimental variation because even modern genomic breeding programs are inefficient in capturing genotype-phenotype relations of low frequency alleles. Our study reveals that the frequency of the haplotypes exhibiting missing homozygosity ranges from 0.5–11%, showing that we have the statistical power to detect very rare deleterious haplotypes in our populations, but also confirming that, as expected, truly lethal recessive variants are invariably infrequent.

The approach chosen for this study relies on the premise that unexpected absence of homozygotes results from unviability of the homozygous deleterious allelic state. Ideally, if we would have sequence data for many thousands of animals, we would be able to directly infer the absence of specific homozygous allele states, e.g. alleles that impair the required protein function, or, alternatively, alleles that are in complete LD with such variants. However, we used low to medium density genotype data, and the SNPs on the chips have been primarily chosen based on their relatively high MAF in most breeds, unlikely to include deleterious variants. To overcome this problem, the haplotype based analysis chosen in this study applied a sliding window approach from 0.5 to 4 Mb. Selecting optimum window sizes is not trivial and depends on population structure, SNP density, recombination rate, and haplotype frequencies within the examined genomic region. For example, by selecting large haplotypes, we increase the risk of analysing recombinant haplotypes. Moreover, by selecting very small haplotypes, non-unique haplotypes might be selected (overlap between distinct haplotypes). We solely used information from complete trios (both parents and offspring genotyped) to calculate the expected number of homozygotes. This number, however, is likely an underestimation, because not all genotyped animals are in genotyped trios. Also, the haplotype approach is unlikely to capture all deleterious variants, as any rare variant that resides only on a common haplotype will be missed. However, rare variants that coincide with rare haplotypes can be robustly detected, with the number of genotyped offspring being the limiting factor for statistical power.

In total, 145 haplotypes showed a significant deficit of homozygotes. Of these, 35 haplotypes showed a negative effect on at least one of the three fertility traits examined, indicating that indeed these 145 haplotypes are highly enriched for variation that can lead to embryonic lethality or prenatal death. The overwhelming majority of these haplotypes are located at chromosomal regions not previously linked to fertility (Hu et al. 2016). Only four genomic loci are shared between the three breeding populations, and all four of these were previously identified as copy number variable (CNV) regions (Paudel et al. 2015). We hypothesize that these four haplotypes are likely false positives, as these CNV events can cause duplication of genetic markers potentially introducing polymorphisms. This could lead to inter-locus cross-hybridization of oligo’s on the chip, causing all individuals to become heterozygous for a particular marker, or a set of markers, generating haplotypes with missing homozygotes.

Several of the identified haplotypes did not show a significant effect on fertility. In some cases, the number of C x C matings was too low to obtain significant statistical support. There are three additional explanations for the absence of an effect on fertility. First, recombination hotspots can potentially result in an excess of heterozygotes that carry recombinant haplotypes. This seems especially apparent at the chromosome ends. Second, since we are examining commercial breeding lines, there is selection on the animals that are genotyped. Piglets are selected for genotyping based on their performance on numerous traits depending on their particular breeding purpose, e.g. growth rate, back fat, fertility, number of teats, and leg quality. Piglets that exhibit unfavourable phenotypes early in life are likely not genotyped and could end up as “missing homozygotes” in our analysis. One example of such a phenotype in pig breeding is the number of teats. Piglets with less than 12 teats are often immediately removed from the population. Third, embryos dying very early (e.g. before time of implantation) are likely replaced by other embryos, as breeding sows are very likely to produce a far higher number of ova than can be accommodated in the uterus.

For one specific haplotype (LW19), identified only in the Large White line, we found evidence from 88 carrier matings that homozygous animals die mid-gestation, and become mummified (fivefold increase in the number of mummified piglets). The association of mummified piglets and haplotype LW19 is very likely an underestimation, since the number of mummified piglets is not always recorded equally strict at all breeding farms, especially for the embryos that died early in gestation for which mummies are small. The haplotype showed a 19% reduction in TNB, less than the 25% loss expected under HWE. One explanation for this discrepancy is that the whole litter might potentially be aborted if a large proportion of the litter dies during gestation, and will therefore not be recorded. We identified the *BMPER* gene as a likely candidate gene causing the defect. In human, the perinatal lethal skeletal disorder diaphanospondylodysostosis (OMIM: 608,022) is associated with homozygous or compound heterozygous mutations in the *BMPER* gene (Funari et al. 2010). Characteristics include a small chest, abnormal vertebral segmentation, and posterior rib gaps (Funari et al. 2010). Homozygous knockout mice exhibit neonatal lethality associated with abnormal lung and skeleton development (Kelley et al. 2009; Helbing et al. 2011). Moreover, heterozygotes for a null allele exhibit abnormal lung development (Kelley et al. 2009). *BMPER* is involved in the negative regulation of bone morphogenetic proteins (BMPs), a group of growth factors involved in the formation of bone and cartilage (Moser et al. 2003). Variation in this gene has been associated with increased body size and rump length in cattle (Zhao et al. 2015), and higher intramuscular fat content in pig (Liu et al. 2014). Evidence for similar early termination of development comes from human and mouse studies (Kelley et al. 2009; Funari et al. 2010). We observed multiple haplotypes with a deficit of homozygotes surrounding the LW19 haplotype associated with similar phenotypic effects (increase in number of mummified piglets, decrease in TNB). It is likely that these haplotypes are not in complete LD with the causal variant. Interestingly, these haplotypes are surrounding a region under selection, despite low LD with the selected haplotype, we hypothesize that this could be a remnant of genetic hitchhiking in the past as this locus has previously been associated with increased body weight and ovulation rate (Hernandez et al. 2014;

Rothammer et al. 2014). Therefore, LW19 might have been subjected to genetic hitchhiking, although we did not find direct evidence supporting this (LW19 is not in LD with the neighbouring haplotypes under selection). More recent recombination might have lowered the LD, but as a result of previous hitchhiking the haplotype still segregates in the population.

Despite the limited impact on crossbred products, given that most haplotypes are population specific, eradication of these haplotypes is still desired. Especially because embryonic lethality leading to mummification does not only have negative economic consequences for the pig breeder, but also results in reduced animal welfare such as health risk for the sow (Christianson 1992). Our study can directly impact positively on current breeding programs, by avoiding C x C matings to lower the frequency of the lethal recessive haplotypes in the elite breeding lines. Furthermore, if a causal variant is found, avoiding homozygotes could be combined with a low-level selection to eradicate the variant after a number of generations. Many risk factors have been associated with an increase in mummified piglets (Le Cozler et al. 2002), most of them, however, are independent of the foetus's genetic material. In our study, we found that about 2% of all recorded mummies in the Large White breed can be attributed to C x C matings for the LW19 haplotype. We believe that the majority of the total recorded mummified piglets are not a direct effect of a genetic defect carried by the unborn foetus, because several other factors, especially many pathogens can cause mummified or stillborn piglets. Therefore, this proportion of 2% likely represents a much larger fraction of the total number of mummified piglets directly caused by a genetic defect carried by the unborn foetus.

The use of molecular tools, in particular in genomic selection, has considerably increased breeding progress over the past years. Despite this, identification of low-frequency deleterious recessive alleles, present in livestock populations, remains a challenging task. The expected low frequency of this type of variation, and therefore marginal effects on fertility traits at the population, contribute to this challenge. However, the routine large scale genotyping of domesticated animals has opened up new possibilities to detect these low frequency deleterious alleles. A systematic genomic survey for missing homozygosity, as applied in this study, is especially promising when thousands of individuals are genotyped. Using this method, novel genetic defects can be identified and fully characterized. Moreover, the increased use of whole genome sequence (WGS) data for breeding purposes opens new opportunities to directly infer deleterious variants from the sequence itself (Charlier et al. 2016).

2.4 Conclusion

Scanning for depletion of homozygous haplotypes provides a powerful tool to identify deleterious recessive alleles, especially for large genotyped livestock populations. Our results confirm the existence, relative rarity, and severe effects on fertility and welfare of lethal haplotypes in commercial livestock populations. We show that these haplotypes, apart from reduced fertility in the parent animal, also cause large numbers (several hundred at population level) of stillbirths ('mummified piglets'), even within highly managed populations. Moreover, the method applied in this study is increasing in potential, as growing numbers of genotyped animals are becoming available for breeding purposes. Finally, this study will facilitate at least partial purging of lethal variation by avoiding matings producing affected or non-viable progeny, demonstrating its value for current breeding programs and animal welfare.

2.5 Methods

2.5.1 Animals, genotypes and pre-processing

The dataset consists of 5517 animals from a commercial synthetic boar line (cross between Large White and Piétrain), and 5301 Landrace and 12,982 Large White animals from two commercial sow lines. Three different SNP panels were used to genotype the animals for the analysis; the 10 K GeneSeek-Neogen Genomic Profiler 10 k BeadChip comprising 10,241 SNP (10 K), the Illumina Infinium PorcineSNP60 v2 BeadChip comprising 61,565 SNP (60 K), and the GeneSeek-Neogen PorcineSNP80 BeadChip comprising 68,528 SNP (80 K). An overview of the number of animals per panel is provided in Additional file 1: Table S1. The chromosomal positions were determined based on the Sus scrofa reference assembly (Groenen et al. 2012). SNPs with unknown position on Sscrofa10.2 and sex-chromosomal SNPs were discarded. Additional file 1: Table S2 provides an overview of the number of SNPs that met the following requirements: Each marker had a MAF greater than 0.01, and a call rate greater than 0.85. Only one marker was used if a genomic position contained multiple markers. Moreover animals with frequency of missing genotypes greater than 0.30 were discarded from the analysis (Additional file 1: Table S4). All pre-processing steps were performed using Plink v1.90b3.30 (Purcell et al. 2007). We did not filter for deviation of HWE because we expect the MH and DH haplotypes to deviate from HWE. The final dataset contained 22,961 animals with an average per-individual call rate of 0.987, 0.955, and 0.983 for 10 K, 60 K and 80 K SNP panels, respectively (Additional file 1: Table S3). Inbreeding assessment was performed by calculating the F coefficient (observed vs expected homozygous genotype counts) in Plink v1.90b3.30 (Purcell et al. 2007).

2.5.2 Imputation from lower to higher density SNP panels

We used the pedigree based BEAGLE version 4.0 genetic analysis software for phasing and imputation of samples (Browning and Browning 2007). First, 10 K samples were imputed to 60 K per individual breeding line. This differs for the boar line in which we imputed 10 K samples directly to 80 K due to the smaller number of 60 K reference samples for this line. Thereafter 60 K was imputed to 80 K and one final round of phasing was performed to make full use of the family relationships (parent offspring duos and trios). Bcftools version 1.3–27-gf31e888 was used to merge vcf files (Li et al. 2009). Imputation accuracies are presented in Additional file 1: Table S3.

2.5.3 Identification of missing homozygote haplotypes

We used a sliding window approach shifting along each chromosome ranging from 0.5 to 4 Mb in steps of 0.5 × window size implemented in a python module. We tested a single haplotype per locus (in case of overlapping haplotypes) with lowest p-value for effect on phenotypes. Haplotypes with a frequency > 0.5% were retained for identification of missing homozygotes. The expected number of homozygotes was estimated using the parental haplotype information with the formula described in Fritz et al., 2013. Moreover, the number of heterozygous offspring from carrier matings was calculated to verify whether there is a deviation from HWE. An exact binomial test was applied to test the number of observed homozygotes with the number of expected homozygotes. Haplotypes were considered significant if $P < 5 \times 10{-}3$ for haplotypes with MH (0 observed) and $P < 5 \times 10{-}6$ for haplotypes exhibiting DH, similar to Pausch et al. 2015. Circos software was used to visualize the haplotypes in genomic ideograms (Krzywinski et al. 2009).

2.5.4 Phenotypic effects associated with lethal haplotypes

Phenotypic records of TNB were available for all three lines. In addition, phenotypic records of NSB, and MUM were available for the two sow lines. In total, records of TNB for 4041 matings comprising 1566 sows and 432 boars were available in the boar line. Records of TNB, NSB, and MUM were available for 15,174 matings comprising 3159 sows and 1485 boars in the Landrace line and 26,961 matings comprising 6745 sows and 1671 boars in the Large White line. We examined each identified haplotype and records on TNB, NSB, and MUM are listed for all C x C matings identified in the phenotypic records. We used a Welch t-test to assess if the phenotypes from the C x C matings significantly differ from C x NC matings. A P-value <= 0.05 was considered significant.

2.5.5 Candidate gene identification

We selected all the genes (Ensembl gene IDs) in regions of missing and deficit homozygosity to perform gene-set enrichment analysis in DAVID (Huang et al. 2007), an enrichment score > 3.0 was considered significant. Also, all porcine genes (Ensembl release 87) within the identified haplotypes were analyzed for the observed phenotypes in gene knock-out/loss-of-function studies in other mammals (mainly for early lethality). Genes that, in knock-out mice, were lethal during developing life stages (embryonic, prenatal, perinatal, neonatal, postnatal or preweaning) were marked as candidate genes (Eppig et al. 2015). Moreover further phenotypic support in other mammalian species was obtained using the OMIM database (Amberger et al. 2015).

2.5.6 Runs of homozygosity and extended haplotype homozygosity

The Large white population was screened for a recent selective sweep at the *BMPER* locus with an EHH test using the R package rehh (Sabeti et al. 2002; Gautier and Vitalis 2012). First, the full dataset was phased with Shapeit v2 (recommended by the rehh package) with inclusion of pedigree information (O'Connell et al. 2014). EHH was generated for each SNP in both populations, identifying long and frequent haplotypes as implemented in the R package rehh (Gautier and Vitalis 2012). The origin and footprint of selection based on haplotype structure was examined using a bifurcation diagram (Sabeti et al. 2002). ROH were inferred using PLINK(v9, (Purcell et al. 2007)) with at least 20 markers covering a ROH, a maximum of one heterozygous call within a stretch and minimum size of 1 Mb.

2.6 Additional files

The online version of this article (10.1186/s12864-017-4278-1) contains supplementary material, which is available to authorized users.

2.7 Acknowledgements

The authors would like to thank Christian Gross, Dick de Ridder, Marcel Reinders, Chiara Bortoluzzi, and Maulik Upadhyay for useful input on this work, and Naomi Duijvesteijn for providing the genotype data.

2.8 Authors’ contributions

MAMG and HJM conceived and designed the study. MFLD and MB performed the data analysis. MFLD wrote the manuscript. HJM, MAMG, MB, BH, and MSL provided useful comments and suggestions and helped to draft the manuscript. Raw data was provided by MSL and BH. All authors read and approved the final manuscript.

2.9 Competing interests

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

3

Early and late feathering in turkey and chicken: same gene but different mutations

Martijn F. L. Derks[1], Juan M. Herrero-Medrano[1], Richard P. M. A. Crooijmans[1], Addie Vereijken[2], Julie A. Long[3], Hendrik-Jan Megens[1], Martien A. M. Groenen[1]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. [2]Hendrix Genetics Turkeys, Technolgy and Service B.V., P.O. Box 114, 5830 AC Boxmeer, The Netherlands. [3]Animal Biosciences and Biotechnology Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, MD 20705 USA

Genetics Selection Evolution (2018) 50(1):7

Abstract

Sex-linked slow (SF) and fast (FF) feathering rates at hatch have been widely used in poultry breeding for autosexing at hatch. In chicken, the sex-linked K (SF) and k+ (FF) alleles are responsible for the feathering rate phenotype. Allele K is dominant and a partial duplication of the prolactin receptor gene has been identified as the causal mutation. Interestingly, some domesticated turkey lines exhibit similar slow- and fast-feathering phenotypes, but the underlying genetic components and causal mutation have never been investigated. In this study, our aim was to investigate the molecular basis of feathering rate at hatch in domestic turkey.

We performed a sequence-based case–control association study and detected a genomic region on chromosome Z, which is statistically associated with rate of feathering at hatch in turkey. We identified a 5-bp frameshift deletion in the prolactin receptor (*PRLR*) gene that is responsible for slow feathering at hatch. All female cases (SF turkeys) were hemizygous for this deletion, while 188 controls (FF turkeys) were hemizygous or homozygous for the reference allele. This frameshift mutation introduces a premature stop codon and six novel amino acids (AA), which results in a truncated PRLR protein that lacks 98 C-terminal AA.

We present the causal mutation for feathering rate in turkey that causes a partial C-terminal loss of the prolactin receptor, and this truncated PRLR protein is strikingly similar to the protein encoded by the slow feathering K allele in chicken.

3.1 Introduction

Sex identification is an important management factor within many commercial livestock operations. In poultry, sexing can be performed by examining feathering rate, a non-intrusive approach to separate males and females at hatch. In turkey and chicken layer breeds, sexing at hatch is crucial for production. However, for broiler breeds this method is mainly applied at the parent stock level. In chicken, the sex-linked dominant K locus, which is located on the Z-chromosome, is responsible for feather development and is associated with delayed emergence of primary and secondary flight feathers (SF), while the k + allele is associated with fast emergence (FF) of flight feathers (Siegel et al. 1957). The status at this locus is widely used for autosexing at hatch (Siegel et al. 1957). Elferink et al. (Elferink et al. 2008) studied the molecular basis of the K allele and identified a 176-kb tandem duplication, which includes part of the genes prolactin receptor (*PRLR*) and sperm flagellar 2 (*SPEF2*) that are associated with the K allele. Moreover, a molecular test was developed to distinguish between homozygous and heterozygous late feathering males (Elferink et al. 2008). The 176-kb duplication causes a 149-amino-acid (AA) C-terminal loss of the PRLR protein and is most likely the causal mutation for the SF phenotype (Bu et al. 2013b). PRLR is a receptor of the anterior pituitary hormone prolactin that belongs to the type I cytokine receptor family and is involved in various physiological processes including many reproductive and developmental processes, such as hair/coat morphology (Bu et al. 2013a). The *PRLR* gene is widely expressed in all embryonic and somatic tissues and its expression is higher in SF than in FF chicks (Luo et al. 2012).

The domesticated turkey (*Meleagris gallopavo*), an important agricultural species and the second largest contributor to world poultry production (FAOSTAT 2017), shows similar SF and FF phenotypes in some commercial lines (Zakrzewska and Savage 1997), which are used for the same selection goal as in chicken, i.e. reliable and easy determination of sex at hatch. The SF phenotype differs between turkey and chicken with SF turkeys generally showing poor feathering even at a later age (Zakrzewska and Savage 1997). Zakrzewska et al. suggested that the dominant sex-linked inhibited feathering (IF) allele K is responsible for the genetic feathering defect in turkey. Interestingly, expression of this defect ranges from almost complete absence of feathers to full feather covering at a later age (> 4 weeks of age), although until 4 weeks of age no apparent differences between SF birds were observed. Moreover, SF turkeys show inferior reproductive efficiency compared with FF turkeys (Renema et al. 2008) and differences in body weight and carcass characteristics (Sikur et al. 2004). The SF phenotype that is under study here differs from a late feathering phenotype that was described in turkey by Asmundson and Abbott (Asmundson and Abbott 1961), which consists in poor feathering at physical maturity (> 20 weeks of age). In chicken, the SF phenotype has been associated with the sex-linked allele K, whereas in turkey, the underlying genetic components and causal mutation have never been investigated. In this study, we used whole-genome sequence data that were obtained from either slow- or fast-feathering turkeys to perform a case–control genome-wide association study (GWAS) for feathering rate at hatch and to investigate its relation to the chicken allele K.

3.2 Methods

3.2.1 Dataset used for sequencing and mapping

We collected blood from 202 animals representing nine commercial turkey lines and that included 12 SF cases and 12 FF cases selected from the same line. For each sample, DNA was extracted and sequenced on the Illumina HiSeq 2000 sequencer, which generated paired-end 101 bp reads. We used the Sickle software to trim sequences (Joshi and Fass 2011), BWA-MEM (version 0.7.15) to map the whole-genome sequencing data to the turkey reference genome (Melgal5) (Dalloul et al. 2010), the Samtools dedup function to remove duplicate reads (Li et al. 2009), the GATK IndelRealigner to perform local realignments of reads around indels (McKenna et al. 2010a) and Qualimap to obtain mapping statistics (Okonechnikov et al. 2016).

3.2.2 Variant detection and post-processing

We performed population-based variant calling using the Freebayes software with the following settings: (1) min-base-quality 10 (to exclude alleles with support base quality < 10), (2) min-alternate-fraction 0.2 (at least 20% of the reads should support the alternate allele in order to evaluate the position), (3) haplotype-length 0 (to avoid generating haplotypes in VCF), (4) ploidy 2 (assuming diploid organism), and (5) min-alternate-count 2 (to have at least two reads that support the alternate allele in order to evaluate position) (Garrison and Marth 2012). Post-processing was performed using bcftools (Li et al. 2009), and variants that were located within 3 bp of an indel, or with a phred quality score and call rate lower than 20 and 0.7, respectively, were removed. The average call rate was about 0.985, and the average transition/transversion (TS/TV) ratio was 2.62, in line with previous findings in turkey (Aslam et al. 2012).

3.2.3 Population statistics

PCA analysis was performed using PLINK (Purcell et al. 2007) on the filtered vcf files and plotted using the default R plotting utilities.

3.2.4 Functional annotation of variants

SnpEffect (Cingolani et al. 2012) was used for variant annotation and the PROVEAN software for variant effect prediction in missense variants. The following variant classes were considered as potential candidate variants: missense, splice acceptor, splice donor, inframe indels, frameshift, stop lost, stop gained, and start lost variants.

3.2.5 Association study and identification of candidate variants

Single locus associations on the genotypes called by freebayes were tested for SNPs and indels in PLINK using permutations to generate uncorrected and corrected p values (Purcell et al. 2007). p values were generated by applying the Fisher's exact test and an adaptive Monte Carlo permutation test was performed with 5000 replications. Variants with a P lower than 1e-5 were considered significant. Manhattan plots were generated using qqman R package (Turner 2014). We selected all significant protein-altering variants and evaluated their putative effect on the protein based on PROVEAN scores and SnpEffect annotations. Moreover, gene ontology (GO) annotations were obtained from the Uniprot database (The UniProt 2017). Phenotype information on *PRLR* null-mutant mice was from Craven et al. (Craven et al. 2001). The ClustalO alignment software (Sievers et al. 2011) was used to align chicken and turkey *PRLR* sequences.

3.2.6 CNV analysis

CNV-seq was used to perform CNV analysis using a log2-threshold of 0.6 and a p value threshold of 0.001 (Xie and Tammi 2009). The optimum window size was automatically computed and ranged from 2.5 to 7.1 kb. The FF sample MG-WUR-121 and the SF sample MG-WUR-136 were used as control samples in CNV-seq analysis for analyses of SF and FF data, both exhibiting average to high coverage (see Additional file 1: Table S1). CNV-seq R utilities were used to plot the CNV events.

3.3 Results

3.3.1 Case–control sequencing and variant detection

To study the molecular mechanisms that underlie feathering rate at hatch in turkey, we selected 12 animals within each group (SF and FF) from one commercial line for whole-genome re-sequencing (WGS) (All female, [see Additional file 1: Table S1]). Moreover, DNA from 178 FF turkeys from various commercial turkey lines was sequenced for additional control samples. The SF turkeys in the population analysed here have phenotypes that are similar to those described for the dominant sex-linked IF allele K by Zakrzewska et al. 1997. Whole-genome DNA was sequenced and resulted in a total amount of 2.17 Tbp (tera base pairs) from 22.48 × 109 paired-end 101 bp reads. Mapping was performed with BWA—mem (version 0.7.15) to the Meleagris gallopavo build 5 (Melgal5) reference genome with an average mappability and coverage of 98.38%, and 10.5×, respectively. We performed population-based variant calling using Freebayes (Garrison and Marth 2012). Next, we filtered out variants with a low-quality (phred quality score < 20) or a call rate lower than 0.7, which resulted in 8,136,213 (post-filtering) variants including 6,595,059 SNPs, and 1,197,170 indels, with an average variant density of 8.4 variants per kb (see Additional file 1: Table S2). We performed PCA analysis on the 24 cases and control animals to assess population stratification; no distinct clustering was observed between the two groups (see Additional file 2: Figure S1).

3.3.2 Functional annotation of variants

We used SnpEff to assign a range of functional classes to the identified variants (Cingolani et al. 2012). The majority of the variants were located in intronic, ncRNA, or intergenic regions (see Additional file 1: Table S3). We identified 231,073 coding (90,370 protein-altering) variants with an overall missense/silent ratio of 0.545, which means that for every two silent mutations (synonymous) one missense mutation is found (see Additional file 1: Table S4).

3.3.3 Genome-wide association study for feathering rate at hatch

The GWAS revealed a significant signal for 134 SNPs on the Z chromosome. None of the detected variants is in perfect LD with the phenotype (see Additional file 1: Table S5). SNPs associated with the SF phenotype are all located on the short arm of the Z chromosome between 7.95 and 9.79 Mb (Figure 3.1) and (see Additional file 2: Figure S2). This region contains 55 protein-coding genes including the *PRLR* and *SPEF2* genes associated with the SF phenotype in chicken.

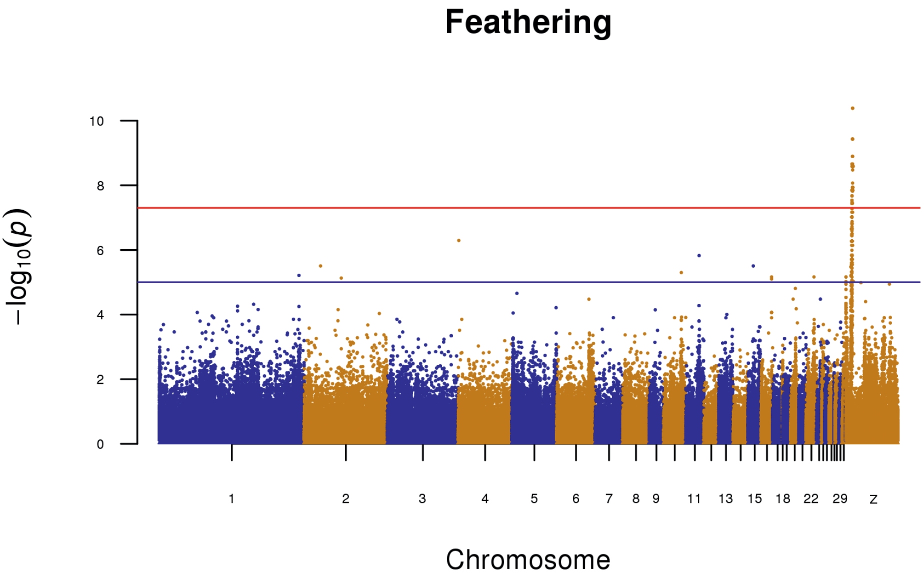


Figure 3.1 Manhattan plot for feathering rate association analysis. The − log10 (P) for each SNP is shown on the y-axis. A clear signal is observed on chromosome Z (8.1–9.5 Mb)

3.3.4 A 5-bp deletion in the *PRLR* gene is associated with slow feathering rate in turkey

We examined the putative effects of all significant variants associated with slow feathering. In total, we identified eight protein-altering variants (seven SNPs and one indel). The seven identified SNPs cause missense mutations in protein coding genes (Table 3.1). None of the missense variants were predicted to have a high impact on the corresponding protein (reaching a PROVEAN score < − 2.5). Moreover, none of the missense mutations were fixed within the group of SF turkeys (Table 3.1), and thus were unlikely to be the causal variant. The identified indel represents a hemizygous 5-bp deletion that is statistically associated with feathering rate and predicted to have a high impact on the *PRLR* gene by causing a frameshift (Figure 3.2). This deletion, which is located within the terminal exon of the *PRLR* gene, produces a truncated PRLR protein by introducing a premature stop codon and adding six novel C-terminal amino acids (DSITET*, Figure 3.2). All SF turkeys were hemizygous for the alternate allele, while ten FF turkey controls and all additional 178 FF turkeys were hemizygous or homozygous for the reference allele (Table 3.2).

Table 3.1 Significant (p < 1e−5) protein-altering variants and predicted impact.

CHR	BP	REF	ALT	P	Case / Control AF	Gene	Type	Effect	Impact (PROVEAN)
Z	7958551	T	C	1.96e-06	0.909 / 0.167	*RGP1*	Missense	Arg227Lys	Neutral (0.58)
Z	7982630	A	G	5.35e-07	0.0833 / 0.833	*CREB3*	Missense	Pro178Ser	Neutral (-1.55)
Z	7982834	A	G	1.96e-06	0.909 / 0.167	*CREB3*	Missense	Val124Ile	Neutral (0.91)
Z	8172555	T	C	9.60e-08	0.0833 / 0.917	*LOC104914814*	Missense	Gln47Arg	Neutral (-0.22)
Z	8181148	A	G	5.35e-07	0.917 / 0.167	*LOC100540309*	Missense	Arg157Lys	Neutral (0.90)
Z	8227879	T	C	5.35e-07	0.0833 / 0.833	*LOC104914815*	Missense	Val320Ile	Neutral (-0.07)
Z	9003502	A	T	2.60e-09	0 / 0.833	*ADAMTS12*	Missense	Leu991Pro	Neutral (2.10)
Z	9426018	G	GTTGGT	2.60e-09	1 / 0.167	*PRLR*	Frameshift	Glu704FS	High

Table 3.2 Genotypes of the 5-bp PRLR deletion for cases (SF) and control (FF) samples.

Group	Phenotype	N	Genotype		
			GTTGGT/GTTGGT or GTTGGT/-	GTTGGT/G	G/G or G/-
Cases	SF	12	0	0	12
Controls	FF	12	10	0	2
Test	FF	178	178	0	0

In addition, we performed a copy-number variation (CNV) analysis to test whether, as in chicken, a CNV event is associated with feathering rate at hatch. Although one region on chromosome Z between 7.9 and 8.1 Mb harboured copy number variants in various samples, none of them were associated with feathering rate at hatch (see Additional file 3).

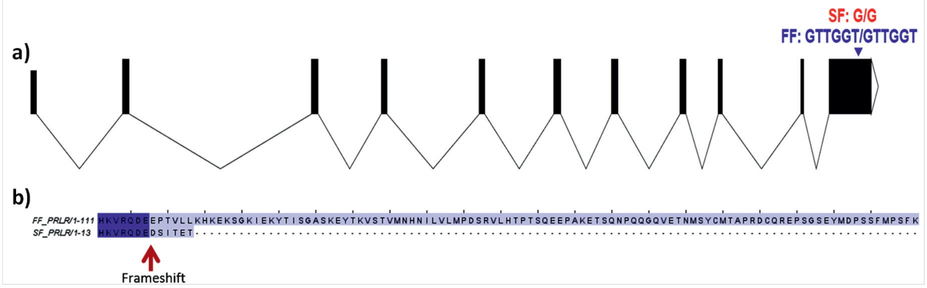

SF: G/G
FF: GTTGGT/GTTGGT
a)
b)
FF_PRLR/1-111 HKVRQDEEPTVLLKHKEKSGKIEKYTISGASKEYTKVSTVMNHNILVLMPDSRVLHTPTSQEEPAKETSQNPQQGQVETNMSYCMTAPRDCQREPSGSEYMDPSSFMPSFK
SF_PRLR/1-13 HKVRQDEDSITET
Frameshift

Figure 3.2 a) *PRLR* gene model showing the location of the 5-bp deletion in the C-terminal exon. b) C-terminal end of the PRLR protein sequence in SF and FF turkey. The deletion associated with SF phenotype introduces a frameshift and six novel AA before a premature stop-codon, which results in the deletion of 98 C-terminal AA.

3.3.5 Chicken and turkey slow feathering

Turkey and chicken PRLR proteins are very similar (90.24% sequence identity, (see Additional file 2: Figure S3) and both are 831 AA long. However, carriers of the 5-bp frameshift deletion in turkey lack the final 98 AA of the PRLR C-end tail (Figure 3.2), whereas carriers of the K allele in chicken lack the terminal 149 AA of the PLRP C-end tail (Bu et al. 2013b). The prolactin receptor forms a dimer to bind prolactin in the extracellular space on the N-terminal end of the protein (Bu et al. 2013a). Moreover, PRLR contains two fibronectin type 3 domains (FN3), a WSXWS motif that is important for proper protein folding, and a Box 1 motif that is necessary for Janus kinase (JAK) interaction and activation (Bu et al. 2013a). However, the function of the affected C-end tail, which is located in the cytoplasm, is mostly unknown, but likely shares similar functional relevance in chicken and turkey.

3.4 Discussion

This study reveals the molecular mechanisms that underlie the rate of feathering at hatch in turkey. The use of NGS data provided a sufficient number of variants to describe the potential causal polymorphism, i.e. a 5-bp deletion within the last exon of the *PRLR* gene. This mutation is different from that of allele K in chicken, but impacts the same gene (Elferink et al. 2008) and moreover, in a similar manner, i.e. by loss of a substantial part of the C-end tail (98 AA in turkey; 149 AA in chicken). Unlike SF chickens, SF turkeys are poorly feathered, even at physical maturity (Zakrzewska and Savage 1997). Moreover, feathering of females can be so poor that carriers of this allele are not used commercially. Although strongly associated with SF, the *PRLR* 5-bp frameshift mutation is not in perfect LD with the phenotype since we observed two FF females that were hemizygous for this deletion. One possibility is that these two animals were mislabeled as FF turkeys, although they are SF turkeys; this is supported by the observation that none of the variants (including non-coding ones) is in complete LD with the phenotype.

The membrane-protein PRLR is a member of the cytokine receptor family that binds the prolactin hormone (PRL) within the extracellular space (Bu et al. 2013a). This hormone is involved in a diverse range of biological activities including various reproductive and developmental processes, such as hair replacement and follicle development (Bu et al. 2013a). Null mutant mice show different hair/coat morphologies and advanced hair replacement (Ormandy et al. 1997). Moreover, a frameshift variant, which introduces a premature stop codon in the bovine PRLR receptor and causes the loss of 120 C-terminal AA, is associated with abnormally short and sleek hair coat (Littlejohn et al. 2014). Moreover, hair development and feather development are considered to have an evolutionary homologous origin. Thus, these findings support the *PRLR* gene as a likely candidate for feathering development within both commercial poultry species, chicken and turkey.

Other studies have suggested that feathering rate in chicken is caused by a higher expression of *PRLR* due to its partial duplication. Carriers of allele K show a 1.78-fold higher expression of *PRLR* in chicken (Luo et al. 2012). In contrast, Zhao et al. found no difference in *PRLR* expression between SF and FF chicks, but that the expression of the other gene involved in the duplication, *SPEF2*, was significantly higher in SF than in FF chicks, which suggested that a mutation in this gene was responsible for the SF phenotype (Zhao et al. 2016). We believe that the higher expression of *SPEF2* in chicken is due to the large duplication that underlies the K allele. The duplication results in two partial *PRLR* genes (that lack both tails), while the *SPEF2* gene remains complete (Elferink et al. 2008). Incomplete *PRLR* mRNA could be subject to the nonsense-mediated decay mechanism resulting in a lower abundance of *PRLR* mRNA compared to *SPEF2* mRNA. Thus, based on our findings, we believe that, rather than a higher expression of *PRLR*, it is the lack of the C-terminal end of the protein that is responsible for the slow feathering rate at hatch in both chicken and turkey. Interestingly, Nakamura et al. reported that, in a late feathering chicken line, reversion to the fast feathering phenotype occurred in rare instances (Nakamura et al. 2011), but this was not observed in our population.

The *PRLR* mutations in chicken and turkey are clearly independent, but lead to similar phenotypes, which strongly suggests that they have been favored by identical breeding goals being applied in these two species. Thus, the SF/FF phenotype shows a pattern that is similar to that observed for a small number of monogenic or oligogenic traits under domestication selection, which show independent mutations in the same genes in specific pathways (Cieslak et al. 2011). Coat colour is one of the most common domestication features, which is regulated by a small number of genes (e.g. *KIT*, *MC1R*, and *TYR*) in many domestic animals (Wright 2015). Another example in poultry is comb morphology, which is a monogenic trait regulated by the same set of genes but with independent mutations in different breeds (e.g. *EOMES*, *MNR2*, and *SOX5*) (Imsland et al. 2012). Thus, we hypothesize that the same independent selection applied for domestic feathering rate within and across species has resulted in independent mutations in *PRLR*.

3.5 Conclusions

We describe a case–control GWAS that detected a genomic region on the Z chromosome, which is statistically associated with rate of feathering at hatch in turkey. Within this genomic region, we identified a hemizygous 5-bp frameshift deletion in *PRLR*, which causes the loss of 98 C-terminal AA and is the causal polymorphism for low feathering phenotype in turkey. This is a clear example of similar selection pressures for the same trait (sexing at hatch) in two domestic poultry species that result in two distinct mutations but each affecting the C-terminal end of the same protein, i.e. PRLR. The function of the C-terminal end of this protein, located in the cytoplasm, remains mostly unknown, and further functional studies are necessary to gain more insight in the downstream molecular pathways affected by this mutation.

3.6 Additional files

The online version of this article (10.1186/s12711-018-0380-3) contains supplementary material, which is available to authorized users.

3.7 Acknowledgements

The authors would like to thank Chiara Bortoluzzi and Vinicius Da Silva for useful input on this work.

3.8 Authors' contributions

MAMG and AV conceived and designed the study. MFLD and JMH performed the data analysis. MFLD wrote the manuscript. JMH, RPMAC, AV, JAL, HJM, and MAMG provided useful comments and suggestions and helped to draft the manuscript. Data collection and sequencing were provided by RPMAC and JAL. All authors read and approved the final manuscript.

3.9 Competing interests

The authors declare that they have no competing interests.

4

A survey of functional genomic variation in domesticated chickens

Martijn F. L. Derks[1], Hendrik-Jan Megens[1], Mirte Bosse[1], Jeroen Visscher[2], Katrijn Peeters[2], Marco C. A. M. Bink[2], Addie Vereijken[2], Christian Gross[3,4], Dick de Ridder[3], Marcel J. T. Reinders[4], Martien A. M. Groenen[1]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. [2]Hendrix Genetics Research Technology & Service B.V., P.O. Box 114, 5830 AC Boxmeer, The Netherlands. [3]Bioinformatics Group, Wageningen University and Research, P.O. Box 633, 6708 PB Wageningen, The Netherlands. [4]Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands.

Genetics Selection Evolution (2018) 50(1):17

Abstract

Deleterious genetic variation can increase in frequency as a result of mutations, genetic drift, and genetic hitchhiking. Although individual effects are often small, the cumulative effect of deleterious genetic variation can impact population fitness substantially. In this study, we examined the genome of commercial purebred chicken lines for deleterious and functional variations, combining genotype and whole-genome sequence data.

We analysed over 22,000 animals that were genotyped on a 60 K SNP chip from four purebred lines (two white egg and two brown egg layer lines) and two crossbred lines. We identified 79 haplotypes that showed a significant deficit in homozygous carriers. This deficit was assumed to stem from haplotypes that potentially harbour lethal recessive variations. To identify potentially deleterious mutations, a catalogue of over 10 million variants was derived from 250 whole-genome sequenced animals from three purebred white-egg layer lines. Out of 4219 putative deleterious variants, 152 mutations were identified that likely induce embryonic lethality in the homozygous state. Inferred deleterious variation showed evidence of purifying selection and deleterious alleles were generally overrepresented in regions of low recombination. Finally, we found evidence that mutations, which were inferred to be evolutionally intolerant, likely have positive effects in commercial chicken populations.

We present a comprehensive genomic perspective on deleterious and functional genetic variation in egg layer breeding lines, which are under intensive selection and characterized by a small effective population size. We show that deleterious variation is subject to purifying selection and that there is a positive relationship between recombination rate and purging efficiency. In addition, multiple putative functional coding variants were discovered in selective sweep regions, which are likely under positive selection. Together, this study provides a unique molecular perspective on functional and deleterious variation in commercial egg-laying chickens, which can enhance current genomic breeding practices to lower the frequency of undesirable variants in the population.

4.1 Introduction

In animal breeding, the number of deleterious genetic variants that are segregating in a population is affected by several factors, e.g. genetic drift, mutation rate, and selection. As a result, small effective population size and artificial selection can impact population fitness in domesticated populations substantially (Charlesworth and Willis 2009) and can lead to a high risk of inbreeding depression, which is the result of the accumulation of deleterious alleles that increase in frequency, mainly due to genetic drift (Charlesworth and Willis 2009). Deleterious alleles are expected to be purged from the population by purifying selection, and thus, generally remain at low frequencies in a population (Zhang et al. 2016b). However, many evolutionary forces shape the landscape of deleterious alleles in a population, including recombination and genetic hitchhiking, which is a change in allele frequency due to the allele being passed along together with a variant that is under selection (Chun and Fay 2011). Recent examples have shown a large impact of such deleterious alleles in several livestock populations (VanRaden et al. 2011; Derks et al. 2017). Therefore, effective purging of these deleterious variants is desired. However, most of these variants are rare, and selection on rare variants is usually inefficient, especially if the relationship between genotype and phenotype is poorly characterized (Kearney et al. 2009; Sitzenstock et al. 2013).

In this study, we examined chicken layer lines that have been primarily selected for production traits, including mortality, egg production, egg composition, shell quality (Ellen et al. 2008), and traits related to animal welfare (Savory 1995). In spite of the many positive consequences of this artificial selection, several health issues are associated with intense selection for production traits in laying hens, including excessive comb growth, brittle bones, feather pecking, and ovarian cancer (Savory 1995; Webster 2004; Johnson et al. 2015). To date, the underlying genetic architecture of these deleterious effects has not been characterized. Therefore, it is essential to better understand the relationship between genotype and phenotype, which is, to a large extent, still a black box (Habier et al. 2013).

Purebred chickens are routinely genotyped by breeding companies using SNP genotyping panels to accelerate genetic progress by applying genomic selection (Meuwissen et al. 2001). Although genomic selection itself may not be very efficient in eliminating low-frequency deleterious variants, the large number of routinely genotyped and pedigreed individuals does allow for the identification of deleterious variation. A powerful method is to systematically assess missing homozygosity in the genome by identifying haplotypes that cause early lethality by statistical depletion, or even absence, of the homozygous state, suggesting that they carry a lethal recessive mutation (VanRaden et al. 2011). This approach can detect even very rare (frequency < 2%) deleterious haplotypes if a large number, at least several thousands, of animals are genotyped in a population. One disadvantage of this method is that low-frequency deleterious variants that reside on common haplotypes will be missed (Derks et al. 2017). An alternative method that does allow such rare deleterious alleles to be identified is to sequence the entire genome of tens to hundreds of animals from a population. Whole-genome sequencing (WGS) can be used to identify potential phenotype-altering variants, which can range from embryonic lethal (EL) to only mildly deleterious mutations in coding regions, and to predict their effects using various tools (Charlier et al. 2016). The use of WGS data from a population can lead to the discovery of variants that are beneficial for breeding programs (Rubin et al. 2010; Gheyas et al. 2015), e.g. by looking for regions in the genome that are under (recent) positive selection. A challenge for this approach is to differentiate true selected variants and variants that increased in frequency as a result of genetic drift. In addition, the incompleteness of current genome annotations in most livestock species hampers the identification of such variants.

In this study, we combined two complementary approaches to identify deleterious and functional variation (positively selected variants in relation to traits under selection) in purebred commercial layer lines. First, we showed that missing homozygosity can result from early embryonic lethality. Second, we mined the genomes of 250 whole-genome-sequenced individuals for deleterious (including embryonic lethal) and functional variants. The result is a comprehensive catalogue of putative deleterious and functional variants, which will be an important resource for future functional studies in chicken and should facilitate the purging of deleterious variants in breeding populations.

4.2 Methods

4.2.1 Animals, genotypes and pre-processing

We genotyped six different commercial chicken breeds using the 60 K Illumina SNP BeadChip: one purebred white layer dam line (WA), one purebred white layer sire line (W1), two crossbred lines (CB: W1-WA, W1-WD) and two brown layer lines (B1, B2) (see Additional file 1: Table S1). All animals from multiple generations were genotyped as part of a routine data collection from Hendrix-Genetics breeding programs. Chromosomal positions were determined based on the *Gallus gallus* GalGal5 reference assembly (Schmid et al. 2015). SNPs with an unknown position on the Galgal5 reference assembly and SNPs on sex chromosomes were discarded. Pre-processing was performed using PLINK v1.90b3.30 (Purcell et al. 2007; Chang et al. 2015) based on the following criteria: each SNP had to have a minor allele frequency higher than 0.01 (1%) and a call rate higher than 0.85 and animals with a call rate lower than 0.7 were discarded from the analysis. We did not filter for deviations from Hardy–Weinberg equilibrium (HWE) because haplotypes that exhibit a deficit in homozygosity were expected to deviate from HWE.

4.2.2 Phasing and identification of missing homozygous haplotypes

We used the BEAGLE version 4.0 genetic analysis software for phasing of the SNP genotypes (Browning and Browning 2007). We used a sliding-window approach using window sizes ranging from 0.25 to 1 Mb in steps of 0.5 times the window size. Haplotypes with a frequency higher than 0.5% were retained for identification of missing homozygotes. The expected number of homozygotes was estimated using the parental haplotype information with the formula described by Fritz et al. (Fritz et al. 2013). The number of heterozygous offspring from carrier matings was also calculated to verify whether there was a deviation from HWE. An exact binomial test was applied to compare the number of observed versus expected homozygotes. Haplotypes were considered significantly depleted of homozygotes if the *p* value for this test was less than 0.005.

4.2.3 Population sequencing and mapping

We used WGS data from three commercial white layer lines, two dam lines (WA: 71, WD: 78) and one sire line (W1: 101), and sequenced a total of 3.502 Tbp (tera base pairs) from 35.94 billion paired-end 100 bp reads sequenced on an Illumina HiSeq machine. We used Sickle software to trim the sequences (Joshi and Fass 2011), BWA-MEM (version 0.7.15, (Li and Durbin 2009)) to map the WGS data to the chicken reference genome (Galgal5) (Schmid et al. 2015), the Samtools dedup function to discard duplicate reads (Li et al. 2009), and GATK IndelRealigner to perform local realignments of reads around indels (McKenna et al. 2010b).

4.2.4 Variant detection and post-processing

We performed population-based variant calling using Freebayes software taking the aligned BAM files as input with the following settings: —min-base-quality 10—min-alternate-fraction 0.2—haplotype-length 0—pooled-continuous—ploidy 2—min-alternate-count 2 (Garrison and Marth 2012). Post-processing was performed using bcftools (Li et al. 2009) and variants that were located within 3 bp of an indel, or with a phred quality score and call rate below 20 and 0.7, respectively, were discarded. Moreover, genotype calls were filtered for sample depth (min: 4, max: AvgDepth * 2.5).

4.2.5 Candidate gene identification

We imputed the 250 WGS animals to 60 K genotypes, to match 60 K-based haplotypes to the available sequence data. The software Confirm-gt (Browning and Browning 2007) was used to match chromosome, strand, and allele to the phased 60 K reference population. BEAGLE version 4.0 was used for imputation and phasing. Carriers of haplotypes that were significantly depleted of homozygotes were examined for causal variants by selecting protein-altering variants carried uniquely by the haplotype carriers. We used the variant effect predictor (VEP, Ensembl-release 86) to predict the impact of the candidate variants identified (McLaren et al. 2016). The impact of the missense variants were assessed using the SIFT and PROVEAN software tools (Kumar et al. 2009; Choi and Chan 2015).

4.2.6 Population statistics

Principle component analysis was performed using PLINK on the filtered vcf files and plotted using the R package ggplot2. PLINK was used with the --het option to calculate the inbreeding coefficient of each individual to assess the level of genetic diversity within each line.

4.2.7 Functional annotation of variants

Annotation of the freebayes-called variants was performed using Variant Effect Predictor (McLaren et al. 2016). Variant effect prediction for protein-altering variants was performed using SIFT (Kumar et al. 2009) and PROVEAN (Choi and Chan 2015). The following variant classes were considered as potentially causing loss of function: splice acceptor, splice donor, inframe indels, frameshift, stop loss, stop gain, and start lost variants. Moreover, only variants that were annotated in genes and which were (mostly) 1:1 orthologous in Ensembl (release 86) were retained to minimize the effect of off-site mapping of sequence reads, as this leads to miscalls, which can be particularly problematic for large gene families (e.g. olfactory receptors). In addition, compensation of function by (recent) paralogous genes will likely ameliorate the effects of damaging mutations in these genes. Also, since gene models might be incorrect, variants that did not have a combined RNA-seq expression coverage of at least 200 in the Ensembl (release 86) merged RNA-seq dataset were discarded. The number and load of deleterious variants for each line were inferred from the final set of deleterious variants.

4.2.8 Spectrum of allele frequencies for different classes of variants

We determined the distribution of allele frequencies for different classes of variants (synonymous, missense tolerated, missense deleterious, stop-gained) to test whether predicted deleterious mutations have generally lower allele frequencies. We generated a histogram with 20 bins (with steps of 0.05 allele frequency) starting from a very low (0–0.05) to very high allele frequency (0.95–1) for the different classes of variants using the PyVCF and SciPy software packages.

4.2.9 Candidate embryonic lethal variants in protein coding genes

To identify putative embryonic lethal (EL) variants, we selected all LoF and deleterious missense variants, for which no individuals that were homozygous for the alternate allele were observed. For every EL candidate we examined whether the gene is known to cause early lethality in mice obtained from the MGI database release 6.10 (i.e. phenotypes from null-mutant mice) (Blake et al. 2017). We manually examined all predicted EL variants in JBrowse (Skinner et al. 2009) to exclude false positives that derived from sequencing and mapping errors. Significant differences in hatchability between carrier by carrier versus carrier by non-carrier phenotypes were assessed using a two-sample t-test, assuming equal variances.

4.2.10 Relative position of indels and stop-gained variants in the protein

We divided proteins from Ensembl release 86 in 10 bins (from N- to C-terminal end) and we determined the relative position of the indel and stop-gained variants by dividing the position of the affected amino acid by the total protein length.

4.2.11 Fixed and line-specific "evolutionary-intolerant" variants

We considered all alleles with a frequency higher than 0.9 (within each line) as fixed or nearly fixed variation. To identify regions under selection, we used an approach similar to that described by Elferink et al. (Elferink et al. 2012), but we applied a window size of 20 kb with a minimum number of 20 variants in each window. We selected a threshold of $zHp \leq -2.7$ representing the extreme lower end of the zHp distribution (see Additional file 2: Figure S1). Windows below this threshold were assumed to be enriched for regions of selective sweeps. We selected line-specific high-frequency variants (i.e. absent in the other two populations) with an allele frequency higher than 0.7.

4.2.12 Gene-set enrichment analysis

We tested whether certain gene families are enriched for deleterious mutations. Therefore, gene-set enrichment analysis was performed using the DAVID functional annotation and classification tools (Huang et al. 2009). Enrichment clusters (as produced by DAVID) with a score higher or equal to 1.3 were considered to be enriched (Huang et al. 2009).

4.2.13 Deleterious alleles in regions of low recombination

The recombination rate is the genetic length in centimorgans divided by the physical genomic distance in mega base pairs and was calculated for bins of approximately 750 kb on macrochromosomes 1 to 5 using the linkage map of Elferink et al. (Elferink et al. 2010). Microchromosomes were excluded because of their extreme high recombination rates (Megens et al. 2009). The ratio of predicted deleterious to predicted tolerated mutations (prediction by SIFT) was calculated within each bin by dividing the number of deleterious missense mutations by the sum of the synonymous and tolerated missense mutations over all three breeding lines. Pearson correlation was used to infer the relationship between the ratio of predicted deleterious to predicted tolerated mutations and the recombination rate.

4.3 Results

4.3.1 Screening for haplotypes that exhibit missing or deficient homozygosity

In layer breeding programs, genetic improvement is primarily achieved on elite purebred lines. These purebred lines are then crossed to produce parent stock production animals that are again crossed to produce the final laying hen production animals, which benefit from the full exploitation of heterosis (Amuzu-Aweh et al. 2015). To successfully screen these purebred lines for missing homozygosity, we assumed that not all deleterious variation has been purged, and that some low-frequency deleterious variation remains in the population. Since we examined carrier by carrier (C × C) matings, 25% of the offspring were expected to be homozygous for the carrier haplotype. In total, we examined six lines for missing homozygosity, one purebred white layer dam line (WA), one purebred white layer sire line (W1), two crossbred lines (CB: W1-WA, W1-WD) and two brown layer lines (B1, B2). In total, information was available for 22,323 (post-filtering) animals genotyped on the Illumina 60 K chicken SNP BeadChip (52,232 SNPs), which provided the statistical power required to detect even very rare haplotypes (see Additional file 1: Table S1). We performed phasing of all data to determine the haplotypes and used an overlapping sliding-window approach to identify haplotypes with a significant deficit in homozygotes.

Table 4.1: Statistics for missing and depleted homozygous SNP haplotypes in four lines of layer chickens.

Lines	WA	W1	CB	B1-B2
Samples	4409	7197	3983	6737
Trios	2291	3619	3539	3118
Number of haplotypes	9	13	7	50
Number of loci	9	13	7	45
Average haplotype length	24.22	33.3	22.29	23.20
Average number of haplotypes per window	17.11	15.08	12.43	15.40
Average haplotype frequency	2.6%	3.1%	8.3%	1.5%
Average homozygous expected	6.06	8.13	30.71	8.08
Average carrier matings with genotyped offspring	3.11	4.23	53.71	3.12
Average carrier matings in pedigree	9.00	12.38	54.71	6.62
Average carrier progeny	24.22	32.54	119.71	32.32
Percentage heterozygote carrier progeny	60.1%	51.3%	70.5%	46.0%
Average number of genes in window	20.9	20.0	9.14	6.30

Averages for all parameters are provided for each line. The number of loci represents the unique number of genomic windows containing significant haplotypes

We identified 9, 13, 7, and 50 haplotypes that exhibited a statistical deficit in homozygosity (DH) in the WA, W1, CB, and B1-B2 lines, respectively (Table 4.1) and (see Additional file 3: Table S1, S2, S3, and S4). The length of these haplotypes ranged from 0.25 to 1 Mb and the frequency of putative deleterious haplotypes ranged from 0.5 to 18.3%. The percentage of heterozygous progeny from C × C matings for these haplotypes was generally higher than 50%, which supports the deviation from HWE due to missing homozygous offspring (Table 4.1). The frequency of these haplotypes was generally low (< 5%) but two haplotypes that showed a deficit in homozygosity had relatively high frequencies (> 10%) in the crossbred line (on *Gallus gallus* chromosome (GGA)1: 180.25–180.75 Mb and GGA5: 5.5–6.0 Mb).

We examined the sequence of the carriers for haplotypes showing a deficit in homozygosity (from the WA and W1 lines) for protein altering variants that were shared by the carriers for each putative deleterious haplotype but for which no homozygous individuals were observed. We identified two candidate mutations (see

Additional file 1: Table S2) that segregated in the purebred (WA and W1) and crossbred lines. These two haplotypes, which were initially identified in the crossbreds (GGA2: 56.0–56.5, GGA3: 94.125–94.875 Mb), contain protein altering mutations in the *ADNP2* (C198S) and *SOX11* (A261G) genes. Both these genes are considered to be essential for normal development and associated with early lethality in mice (inferred from null-mutants, (Pinhasov et al. 2003; Jiang et al. 2013)). Only the alanine to glycine mutation in the *SOX11* genes was predicted to be mildly deleterious by SIFT and PROVEAN (see Additional file 1: Table S2).

4.3.2 A catalogue of genomic variation in three white-layer lines

We also explored the use of WGS data for direct inference of deleterious variation using sequence data from three commercial white layer lines, one sire line (W1), and two dam lines (WA and WD). We sequenced 250 animals from these lines (WA: 71, WD: 78, and W1: 101), for a total volume of 3.502 Tbp (tera base pairs) from 35.94 billion paired-end 100 bp reads. Mapping was performed with BWA-MEM (version 0.7.15, (Li and Durbin 2009)) to the *Gallus gallus* build 5 reference genome (Schmid et al. 2015) with an average mappability and coverage of 99.76%, and 11.4 (range: 8.3X to 22.9X), respectively (Pipeline overview [see Additional file 2: Figure S2]). We performed population-based variant calling using Freebayes (Garrison and Marth 2012) to identify 10,260,277 (post-filtering) variants in the three lines (see Additional file 1: Table S3). From the total 10,260,277 (post-filtering) identified variants, 9,469,408 (98.5% biallelic) were SNPs and 790,869 were indels. The average SNP density was 11.0 per kb (see Additional file 1: Table S3). We identified 2,143,367 novel variants (20.89%) that were not annotated in dbSNP (build 147), of which the majority was breeding line specific (WA, WD, or W1) (see Additional file 1: Table S4). An average call rate of 0.95 and an average transition/transversion (TS/TV) ratio of 2.53 were found for the entire variant set (see Additional file 2: Figure S3 and Additional file 1: Table S5), which are congruent with previous findings in other avian species (Aslam et al. 2012; Smeds et al. 2016). Sample origin was validated using principal component analysis (PCA) (see Additional file 2: Figure S4).

We assessed the level of genetic diversity by calculating the F statistic within the three lines (WA, WD, and W1) and observed that it was lower in the WA line than in the other two lines (see Additional file 2: Figure S5). Accordingly, we found a smaller number of line-specific SNPs in the WA line compared to the other two lines (see Additional file 1: Table S4). Moreover, we observed that WA animals carried on average fewer deleterious variants than the other two lines. However, the mutation load, calculated as the ratio of deleterious (SIFT < 0.01) to synonymous variants, was higher in the WA line than in the WD and W1 lines, which was in line with the lower genetic diversity within this line (Figure 4.1).

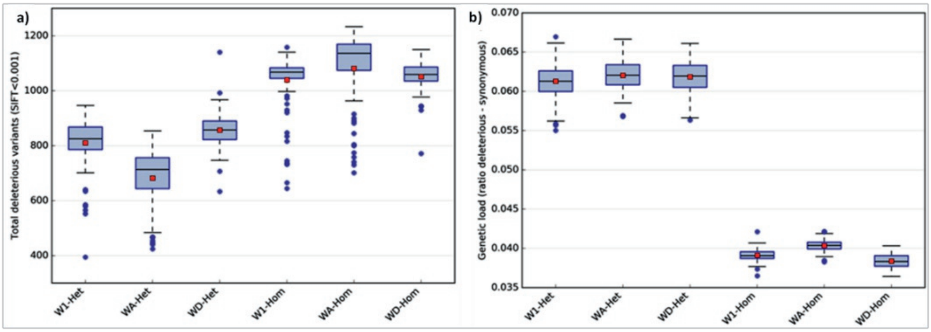


Figure 4.1 a) Distribution of the number of heterozygous (-Het) and homozygous (-Hom) individuals for putative deleterious variants. b) Mutation load, calculated as the ratio of deleterious to synonymous variants for heterozygous and homozygous individuals for putative deleterious variants.

Variant effect prediction assigned a range of functional classes to the identified variants (see Additional file 1: Table S6). Of the 120,149 coding (35,963 protein-altering) variants that we identified, the large majority were synonymous and non-synonymous mutations. Furthermore, 2.04% (2437) of the variants were classified as potentially introducing a loss-of-function (frameshift, inframe deletion, inframe insertion, splice acceptor, splice donor, start lost, stop gained, and stop lost variants). Of the 33,492 missense mutations, 5546 and 3053 were predicted to be deleterious by the SIFT and PROVEAN software, respectively, of which 1847 were predicted by both methods (see Additional file 2: Figure S6). A final set of 4219 putative deleterious variants, distributed across nine classes of deleterious variants, was obtained after filtering (see "Methods") and (see Additional file 1: Table S7).

4.3.3 Evidence for purifying selection on deleterious mutations

We found that the spectrum of allele frequencies of deleterious variants differed from that of neutral variants, and was skewed towards a higher proportion of low-frequency alleles (Figure 4.2) and (see Additional file 2: Figure S7). Their relative low frequency supports the hypothesis that the predicted deleterious variants are subject to purifying selection.

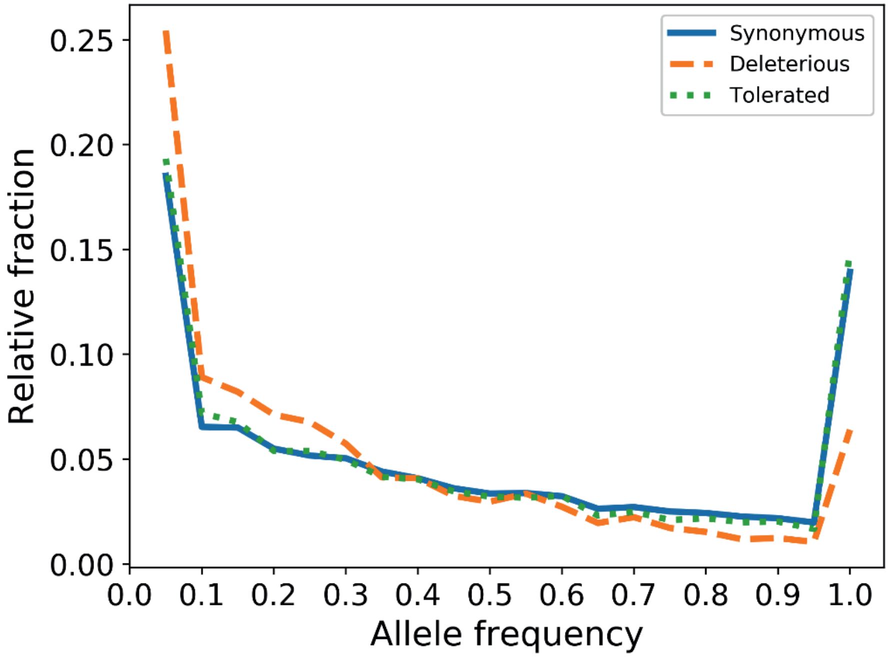


Figure 4.2 Allele frequency distribution for different functional classes of putative deleterious variants. Deleterious variants (deleterious missense and stop-gained) show distinct allele frequency spectra compared to variants considered to be neutral (synonymous, missense tolerated). Missense variants are classified by SIFT (deleterious: SIFT score ≤ 0.05, tolerated: SIFT score > 0.05).

4.3.4 Relative position of indels and stop-gained variants in the protein

The impact of LoF variants on the protein is potentially determined by the position of the variant in the amino acid sequence. We found that frameshift and stop-gained variants were enriched at the N- and C-terminal ends of the protein, a pattern that was not present for inframe indels, which rather showed a more or less uniform distribution of location across the protein (Figure 4.3a). Frameshift or stop-gained variants at the N-terminus could be “rescued” by alternate start-codons, while variants at the C terminus are less likely to be disruptive because they may still result in a more-or-less functional protein. Moreover, deleterious missense mutations occurred more often at the N- and C-terminal ends of the protein, while synonymous mutations occurred less frequently at those positions (see Additional file 2: Figures S8 and S9). Overall, coding indels were enriched for in-frame indels (e.g. 3, 6, 9 bp), because these are more likely to be evolutionary-tolerated (and therefore not purged from the population), which usually does not apply to frameshift indels (Figure 4.3b).

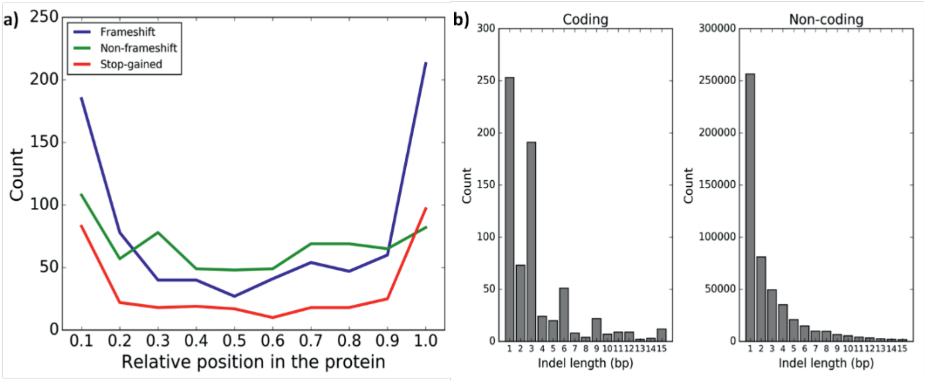


Figure 4.3 a) Relative position of frameshift, non-frameshift indels, and stop-gained variants. Frameshift variants are enriched in N- and C-terminal parts of the protein. Frameshift variants at the N-terminal sites are potentially "rescued" by alternate start-codons. Frameshift variants at the C terminal end are likely not disruptive since a functional protein might still be translated. b) Distribution of lengths of coding and non-coding indels. In-frame indels (i.e. indels with lengths of 3, 6, and 9 nucleotides) are enriched in coding regions.

4.3.5 Less effective purging in regions of low recombination

Next, we examined whether the ratio of deleterious to tolerated mutations was affected by the recombination rate. A significant negative correlation (r = − 0.26, $p = 2.89\times10e^{-9}$) was found between the recombination rate and the ratio of deleterious to tolerated alleles, providing evidence of more effective purging in regions with high recombination rates (Figure 4.4). Enrichment of deleterious over tolerated variants was especially evident in regions of very low recombination (recombination rate less than 2%, [see Additional file 2: Figure S10]).

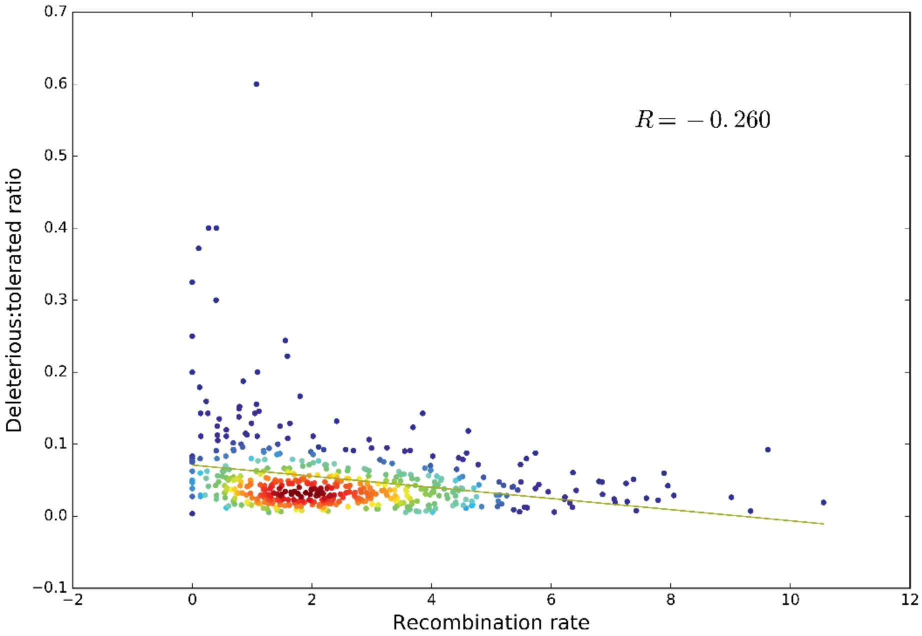


Figure 4.4 Pearson correlation between recombination rate and the ratio of putative deleterious to tolerated alleles for regions that harbour such alleles. Results indicate that regions of low recombination are generally enriched for deleterious variants (R = − 0.26, P = 2.89e−09)

4.3.6 Candidate EL variants in protein coding genes

To identify variants that likely result in early lethality during development (EL), we selected all putative LoF and deleterious missense variants that met the following two criteria: (1) no homozygous individuals for the allele were observed; and (2) the affected gene caused early lethality in null-mutant mice (Blake et al. 2017). Based on these criteria, we identified 11 frameshift, five inframe indels (predicted as deleterious by PROVEAN), six stop-gained, five splice acceptor, eight splice donor, and 121 deleterious missense variants (see Additional file 4: Table S1). The majority of these 152 candidate EL variants (86.6%) were specific to one line and contained frameshift mutations in the *APAF1* and *NHLRC2* genes, which are both associated with embryonic lethality and malformations in cattle (Denholm 2015; Adams et al. 2016). Of the five in-frame indels, two exhibited relatively high carrier frequencies (> 5%) in the WD line and affected the genes *CHTF18* and *FLT4*. We also identified 13 candidate splice donor and acceptor variants that could potentially lead to mis-splicing, resulting in an incomplete or incorrect protein. Two splice variants exhibited relatively high allele frequencies (> 5%) and affected the *POLR1B* and *HP1BP3* genes. Moreover, one high-frequency (22.3%) stop-gained variant affected the C-terminal end of the SCRIB protein and, thus, might not be disruptive as an almost complete functional protein should be translated (see Additional file 4: Table S1).

4.3.7 Missense variants

The large majority (~ 84%) of the 122 candidate EL missense variants were specific to a line (WA: 19, WD: 46, and W1: 37). Twenty-five variants were predicted to be highly deleterious (PROVEAN score < − 5, Table 4.2, and [see Additional file 2: Figure S11]). One specific missense variant in the *OFD1* gene, which causes a tyrosine to cysteine substitution (Y19G), is a strong candidate for embryonic lethality in homozygous carriers, in spite of its relative high frequency (8.9%). The tyrosine at position 19 of OFD1 is highly conserved among vertebrates and, thus, this missense mutation is predicted to be highly deleterious (PROVEAN: − 7.42, SIFT: 0.0). From the 18 carrier animals (15 sires and 3 dams), we identified three C × C matings in the breeding data that showed a significant (*p* = 0.0165) increase in the percentage of embryos that died during development (see Additional file 1: Table S8).

Table 4.2 Missense variants predicted to be highly deleterious (PROVEAN score < − 5.0) and their phenotypic consequences in null mutant mice based on the MGI database.

				#			AA	AA	SIFT		MGI Phenotype
Chr.	Position	Ref.	Alt.	Het.	Line	Symbol	position	change	score	Provean	
1	49464407	C	T	7	W1	NAGA	271	R/C	0	-6.42	Homeostasis
1	51524116	A	G	3	WA	TMPRSS6	688	E/G	0	-5.771	Reproductive, growth/size/body, endocrine/exocrine, liver/biliary, immune, homeostasis, mortality/aging, integument, hematopoietic, digestive/alimentary
1	118357796	A	C	7	W1	PRDX4	81	F/V	0.02	-5.65	Reproductive, cellular, endocrine/exocrine
1	122963712	T	C	18	W1	OFD1	19	Y/C	0	-7.42	Embryo, nervous, system, skeleton, craniofacial, limbs/digits/tail, renal/urinary, respiratory, cellular, mortality/aging, cardiovascular, growth/size/body, digestive/alimentary
2	66538263	T	A	9	WA	BPHL	70	D/V	0	-8.815	Hearing/vestibular/ear, homeostasis
3	108278046	C	T	3	WD	PKHD1	3397	R/W	0.03	-5.066	Respiratory, growth/size/body, endocrine/exocrine, liver/biliary, renal/urinary, cellular, mortality/aging, cardiovascular, nervous system, hematopoietic, digestive/alimentary
4	51931175	G	T	4	W1	CENPC	729	P/Q	0	-7.783	Embryo, mortality/aging, growth/size/body, cellular
4	62191743	T	C	3	WD	FAT1	796	Y/C	0.01	-7.534	Nervous system, craniofacial, renal/urinary, vision/eye, mortality/aging, pigmentation, growth/size/body, homeostasis
4	70191405	G	A	15	WD	TBC1D1	182	R/C	0	-6.635	Growth/size/body, adipose, cellular, no abnormal phenotype observed, muscle, homeostasis
5	58052925	G	A	7	W1	NIN	1206	R/C	0.02	-6.045	Hearing/vestibular/ear, nervous system, behaviour, cardiovascular
5	58235842	C	T	8	W1	NID2	944	G/S	0	-5.151	Immune, skeleton
9	17236801	C	A	3	WA	CCDC39	857	P/H	0	-5.599	Respiratory, skeleton, craniofacial, liver/biliary, immune, renal/urinary, homeostasis, cellular, mortality/aging, digestive/alimentary, growth/size/body, hematopoietic, cardiovascular
9	17571682	C	A	3	WD	MFN1	439	G/W	0	-7.551	Embryo, mortality/aging, growth/size/body, cellular
19	4443604	A	T	11	WD	UNC45B	225	I/N	0	-5.4	Mortality/aging
19	6266804	C	T	9	WD	CPD	994	P/L	0.01	-6.728	Respiratory, behaviour, reproductive, craniofacial, endocrine/exocrine, liver/biliary, immune, digestive/alimentary, homeostasis, cellular, vision/eye, integument, nervous system, skeleton, growth/size/body, hematopoietic, cardiovascular
24	4359027	G	A	3	WA	KMT2A	1941	P/L	0	-9.253	Embryo, liver/biliary, muscle, cellular, reproductive, immune, craniofacial, limbs/digits/tail, hearing/vestibular/ear, renal/urinary, neoplasm, homeostasis, behaviour, cardiovascular, mortality/aging, integument, nervous system, growth/size/body, hematopoietic, skeleton
27	3474503	C	A	4	WA	MPP3	206	S/Y	0	-5.649	Nervous system, vision/eye, cellular

4.3.8 Fixed evolutionary-intolerant variants include potential selection candidates

We identified 473 predicted deleterious alleles that were fixed (247) or nearly fixed (allele frequency > 90%) in the three white layer lines (WA, WD, and W1) (see Additional file 5: Table S1). Gene-set enrichment analysis showed that the corresponding genes are involved in energy metabolism (e.g. ATP-binding, calmodium-binding) and muscle and motor activity (see Additional file 5: Table S2). Several of these variants were strongly selected in domesticated chicken. For example, variant (G558R) in the *TSHR* gene was completely fixed in all three lines and this mutant allele is associated with the absence of strict regulation of seasonal reproduction found in natural populations (Rubin et al. 2012). A deleterious inframe deletion (108delE) was also found in the *P2RY2* gene, which is an ATP receptor. In addition, 12 fixed deleterious variants were identified in seven myosin-related genes (*MYH7B*, *MYCBPAP*, *MYO1G*, *MYH9*, *MYLK3*, *MYO9B*, and *MYLK2*) that are involved in skeletal muscle development (Lagrutta et al. 1989). Other gene families that contained fixed deleterious variants were the protein-tyrosine-phosphatases (*PTPN7*, *PTPRJ*, *TNS3*, *PTPRE*, *PTPRF*, and *DUSP28*), the centrosome proteins (*CEP97*, *CEP162*, *CEP89*, and *CEP164*), which are potentially involved in essential developmental processes, based on evidence of early lethality in knockout model organisms (notably *CEP97* and *CEP164,* Blake et al. 2017), and collagen-like genes (e.g. *C1QTNF8*, *C1QTNF6*, *EMILIN2*). Forty variants in 37 genes were predicted to have a severe impact on the protein produced by these genes (PROVEAN score ≤ − 5), including a variant in the *TSHR* gene (see Additional file 5: Table S3).

4.3.9 Selection candidates

To distinguish between true selection candidates and effects of genetic drift, we examined the populations for regions under selection. Genome-wide Z-scores of heterozygosity (zHp) were calculated per 20-kb windows. We considered bins with a zHp less than − 2.7 as potential regions of selective sweeps in the genome (representing the extreme end of the distribution) (see Additional file 2: Figure S1) and found 27 fixed evolutionary intolerant variants in these regions (see Additional file 2: Figure S12 and Additional file 5: Table S4), which overlap with the *TSHR* (see Additional file 2: Figure S13) and *FOXI1* genes, previously described as being under domestication selection (Rubin et al. 2010; Gheyas et al. 2015).

We focused on predicted evolutionary-intolerant variants in smaller regions of selective sweeps to identify possible functional variation that has been under selection. We identified a splice donor variant in the *CPE* gene (see Additional file 2:

Figure S14), which is involved in the energy metabolism of cells and insulin processing. In addition, we identified a strong selection signal in two bins that overlapped with a missense variant in the *CCDC93* gene (T389 M) (see Additional file 2: Figure S15). This gene is involved in protein transport, but, although various quantitative trait loci (QTL) related to egg production and egg quality overlap with this gene (Hu et al. 2016), its exact function remains unknown. A splice acceptor variant in the *PSMC6*gene, a start lost variant in the *GLCCI1* gene, and an inframe insertion in the *RUNXT1* gene were identified as potential additional functional target mutations (see Additional file 2: Figures S16, S17 and S18). *PSMC6* and *GLCCI1* are both involved in energy metabolism, and overlap respectively with an egg shell thickness QTL and a QTL for haugh unit (a measure of egg protein quality based on the height of its egg white) and growth (Hu et al. 2016). The *RUNX1T1* gene is a transcription factor involved in the generation of precursor metabolites (substances from which energy is derived). All these variants are likely functional, and while they are identified as being damaging in a natural or wild context, they may have been favourably selected for because they positively affect desired traits in egg-laying hens.

4.4.10 Line-specific high-frequency deleterious variation

4.4.10.1 WA breeding line

We found 26 high-frequency (allele frequency > 0.7) deleterious missense variants, one frameshift and three splice variants specific to the WA breeding line. Interestingly, the *ASPM* gene contains three deleterious missense variants (see Additional file 6: Table S1). This gene encodes a mitotic spindle protein and is expressed in proliferating tissues and is associated with a range of phenotypes, including decreased body weight, microcephaly, and reduced fertility in both sexes. Two variants were predicted to have a severe impact on CIB1 (R112C) and PCSK6 (R87 W) proteins (PROVEAN score < − 5), which are both involved in mammalian fertility. CIB1 is related to abnormal spermatogenesis, decreased testis weight and male infertility, while PCSK6 showed a role in female fertility (ovary cysts, increased ovary tumour incidence) (Blake et al. 2017).

4.4.10.2 WD breeding line

We annotated 77 high-frequency deleterious variants specific to the WD breeding line (see Additional file 6: Table S2), which included 59 deleterious missense variants, one inframe deletion (*ENSGALG00000030853*), 14 splice acceptor/donor variants, one start-loss variant (*PCBD2*), and two stop-gained variants (*BRIC5* and *NCOR1*).

Interestingly, the *FYCO1* gene, which is associated with cataract phenotypes in mammals (Blake et al. 2017), harbours two highly deleterious missense variants. Moreover, six missense variants are predicted to be highly deleterious by PROVEAN (*PIGX*, *CARMIL2*, *LPAR6*, *ENSGALG00000015226*, *LIMK2*, *RIC3*). Three of these genes were demonstrated to have severe effects in null-mutant mice (*CARMIL2*, *LPAR6*, and *LIMK2*) (Blake et al. 2017).

4.4.10.3 W1 breeding line

We identified 35 high-frequency variants specific to the W1 breeding line (see Additional file 6: Table S3), which included 31 deleterious missense variants, three splice-donor variants, and one stop-gained variant (*NOLC1*). Three missense variants in three different genes (*TAAR1*: Y290 N, *VWA1*: P251S, *MCM10:*P39L) were predicted to be highly deleterious. *TAAR1,* a trace amine associated receptor gene, and *VWA1* are both associated with various behavioural traits, including increased hyperactivity (*TAAR1*) and abnormal motor coordination/balance (*VWA1*). Null-mutants for the *MCM10* gene are embryonic lethal in mammals, resulting in abnormal growth prior to termination of development (Blake et al. 2017). Interestingly, the *CSPG4*gene harbours three deleterious missense variants in the W1 line, which are associated with abnormal muscle cell physiology and increased body weight (Grako et al. 1999).

4.4 Discussion

Combining a systematic genomic survey for missing homozygosity and whole-genome sequence (WGS) data opens new opportunities to directly infer functional variants. We have presented a first full genomic catalogue of variants that provides a perspective on the deleterious and functional variation in fairly closed, and relatively inbred, purebred layer lines. We not only confirmed previous “domestic” or selective variants but also assessed the impact of deleterious variation in these lines. Taken together, this genomic framework can be used to further improve and understand the genomic elements that are selected or purged in current breeding programs. Finally, a better understanding of the variants with functional implications will provide a useful resource for further selection programs to help distinguish true deleterious variants from those with positive functional implications.

Domesticated populations are expected to be under artificial selection against inbreeding depression. Indeed, in this paper, we show that putatively highly deleterious (i.e. lethal) variants are rare in the commercial chicken populations studied here, in spite of the small effective size of these populations. However, we found several examples of putative lethal variants with allele frequencies up to 10% (e.g. OFD1 and Y19C) and showed that, although under strong selection, the purging of these variants is not always very effective, even in modern poultry breeding programs. Artificial selection in these populations may be ‘strong’, but is based on an index of a large number of phenotypic traits. Balancing selection may also be acting on these populations (e.g. heterozygote advantage), which causes deleterious variants to remain in the population.

In order to capture deleterious variants using haplotypes of SNPs that exhibit missing homozygosity, the low-frequency haplotype has to be in complete LD with the causal variant. However, most deleterious variants (EL) reside on common haplotypes that cannot be detected with medium-density SNP chip data. However, absence of specific homozygous allele states can now be inferred directly because animals can be routinely genotyped for these variants, such that they can be added to the currently used genomic selection framework. A similar study in cattle showed that 15% of the LoF and 6% of the tested missense variants are likely true EL (Charlier et al. 2016). Although predicting EL variation from sequence can be sensitive to induce false positives, we tried to reduce the number of false positives by manually examining the predicted EL variants. Moreover, the distinct allele frequency spectrum for our predicted deleterious mutations compared to neutral mutations confirms that they are subject to purifying selection.

One limitation of our study is that we focused on coding variation, however, a large proportion of the non-coding genome is also subject to purifying selection because of their biological function (Ponting and Hardison 2011). As a result, we may have missed a large proportion of potential deleterious or functional variants. In addition, livestock genomes still lack proper annotation of many functional elements but currently there are many efforts to improve this aspect (Andersson et al. 2015).

We found no evidence of a higher load of deleterious variants in our studied chicken lines compared to other livestock species (Bosse et al. 2015; Charlier et al. 2016). However, although the impact of individual variants on the population may be limited, a recent study showed that negative selection involves synergistic epistasis, which means that the combined effect of mutations is greater than the sum of the individual effects. This supports the hypothesis that the overall effect of the deleterious mutations on population fitness might be substantial (Sohail et al. 2017). As a consequence, the number of deleterious variants found in the chicken populations studied here might represent a universal level for 'healthy populations', i.e. lower levels deleterious mutations are not attained because selection against low-frequency alleles is ineffective, but higher levels of deleterious mutations could occur, which then rapidly leads to disproportionately large inbreeding depression effects. This study also demonstrates the value of domesticated populations to provide insight in the genomic architecture of inbreeding depression and can be useful for future studies on inbreeding in both wild and domesticated populations.

The observed spectrum of allele frequencies for predicted deleterious and tolerated variants corroborates the hypothesis that the predicted deleterious variants (especially deleterious missense and stop-gained variants) have been under purifying selection. Conversely, the predicted tolerated missense variants followed the same distribution of allele frequencies as synonymous variants (usually considered to be neutral), which indicates that the large majority of these predicted missense variants are indeed evolutionary tolerated. Within coding regions, we also found an enrichment of indels that are multiples of three nucleotides, which was not the case for non-coding indels. Indels that alter the frame of translation in coding regions can be highly disruptive, for instance by introducing a premature stop codon and, therefore, such indels are often under purifying selection. Conversely, indels that are multiples of three nucleotides will result in losses or gains of one or multiple amino acid residues, which have a higher likelihood of being tolerated. We also observed an enrichment of frameshift and stop-gained variants at the N- and C terminal ends of the protein, which suggests that, in general, these types of variants have a stronger impact on the function of the protein when they are located in the middle part of the protein compared to the distal parts of the protein. Namely, if they are located at the N-terminal part of the protein, a functional protein product might still be generated by an alternate start codon that can “rescue” a large part of the protein (N-terminal part), as described previously (Ng et al. 2008). In contrast, a frameshift or stop-gained variant at the C-terminal end may be tolerated since an almost complete protein is often generated. Together these genomic signatures of purifying selection support our predictions on deleterious alleles within the populations.

Evidence that the frequency of recombination in a genomic region is negatively correlated with the ratio of deleterious to tolerated mutations suggests more effective purging in regions with higher recombination rate, potentially because deleterious variants that hitchhike along with selected variants are more easily physically disconnected from variants that are under selection in regions with high recombination rates. Similar results have been reported in other species, although always with weaker correlations (Chun and Fay 2011; Zhang et al. 2016b; Ramu et al. 2017). We shed light on the role of recombination (i.e. more effective selection in regions of high recombination) in genomic purging within the avian clade, which is known for its highly diverse recombination rates between chromosomes, with notably extremely high recombination rates on microchromosomes (Backstrom et al. 2010).

In addition to predicted deleterious variants with low frequencies, several high-frequency predicted deleterious variants were identified that likely have high functional relevance. We focused on predicted evolutionary-intolerant, but high-frequency, variants in selective sweep regions. This study confirmed several predicted deleterious variants that were previously identified as being positively selected in domesticated chicken populations, e.g. variants in the *TSHR* and *FOXI1L* genes (Rubin et al. 2010; Gheyas et al. 2015). However, we find several novel predicted deleterious variants in strong selective sweep regions (e.g. variants in the *CCDC93*, *PSMC6* and *GLCCI1* genes), that should be further investigated for phenotypic effects. In spite of a paucity of functional annotation, there is evidence that the majority of these genes have a role in cellular energy metabolism and likely cause increased metabolic activity (Rubin et al. 2010; Elferink et al. 2012).

The use of genomic selection has increased the rate of genetic improvement in breeding populations substantially over the past years (Sitzenstock et al. 2013). However, genomic selection remains a "black-box" approach and the genomic architecture that underlies selection remains unknown. Without additional prior information on the functional effects of low-frequency variants, effective selection for or against desired or unwanted variation remains challenging. Leveraging low-frequency functional variants for selection requires functional annotation, which can then be translated into statistical priors in enhanced genomic selection programs (Perez-Enciso et al. 2015; MacLeod et al. 2016; Perez-Enciso et al. 2017). This study contributes to this by the identification of specific variants that can be incorporated in breeding programs to enhance genetic improvement.

4.5 Conclusions

In this study, we applied several methods to infer deleterious variation in three commercial white-layer lines. We confirmed that missing homozygosity can result from lethal variants that reside on low-frequency SNP haplotypes. We were able to capture even very low-frequency deleterious variation, including 152 likely EL variants, by exploiting WGS data of dozens of sequenced individuals within single populations. Results provided clear evidence for purifying selection, based on a distinct spectrum of allele frequencies of deleterious variants compared to that of variants that have a higher likelihood of being neutral. In spite of their low-frequency nature, the identified putative deleterious alleles generally occurred more often in regions with low recombination, which suggests that purging of these alleles is less effective in such regions. Also, frameshift and stop-gained variants were more frequent at the protein N- and C-termini, which confirms that these are likely evolutionary-tolerated, which also applies to in-frame indels. In addition, multiple predicted evolutionary intolerant coding variants were discovered in selective sweep regions, which are likely under positive selection. A comprehensive genomic catalogue of putative deleterious variants was developed for white-egg layer breeding lines, which can enhance current genomic breeding practices to lower the frequency of undesirable variants in the population.

4.6 Additional files

The online version of this article (10.1186/s12711-018-0390-1) contains supplementary material, which is available to authorized users.

4.7 Acknowledgements

The authors would like to thank Chiara Bortoluzzi, and Vinicius da Silva for useful input on this work.

4.8 Authors' contributions

MAMG and HJM conceived and designed the study. MFLD performed the data analysis and wrote the manuscript. HJM, MAMG, MB, AV, MCAMB, KP, JV, CG, DdR, and MJTR provided useful comments and suggestions and helped to draft the manuscript. AV, MCAMB, KP, and JV provided raw data. All authors read and approved the final manuscript.

4.9 Competing interests

MCAMB, JV, KP, and AV are employees of Hendrix-Genetics, one of the funders of this study. All authors declare that the results are presented in full and as such present no conflict of interest. The other Breed4Food partners Cobb Europe, CRV, Topigs Norsvin, declare to have no competing interests for this study.

5

Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome

Martijn F. L. Derks[1], Marcos S. Lopes[2,3], Mirte Bosse[1], Ole Madsen[1], Bert Dibbits[1], Barbara Harlizius[2], Martien A. M. Groenen[1], Hendrik-Jan Megens[1]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. [2]Topigs Norsvin Research Center, Beuningen, the Netherlands. [3]Topigs Norsvin, Curitiba, Brazil

PLoS genetics (2018) 14 (9), e1007661

Abstract

Livestock populations can be used to study recessive defects caused by deleterious alleles. The frequency of deleterious alleles including recessive lethal alleles can stay at high or moderate frequency within a population, especially if recessive lethal alleles exhibit an advantage for favourable traits in heterozygotes. In this study, we report such a recessive lethal deletion of 212kb (del) within the BBS9 gene in a breeding population of pigs. The deletion produces a truncated BBS9 protein expected to cause a complete loss-of-function, and we find a reduction of approximately 20% on the total number of piglets born from carrier by carrier matings. Homozygous del/del animals die mid- to late-gestation, as observed from high increase in numbers of mummified piglets resulting from carrier-by-carrier crosses. The moderate 10.8% carrier frequency (5.4% allele frequency) in this pig population suggests an advantage on a favourable trait in heterozygotes. Indeed, heterozygous carriers exhibit increased growth rate, an important selection trait in pig breeding. Increased growth and appetite together with a lower birth weight for carriers of the BBS9 null allele in pigs is analogous to the phenotype described in human and mouse for (naturally occurring) BBS9 null-mutants. We show that fetal death, however, is induced by reduced expression of the downstream BMPER gene, an essential gene for normal foetal development. In conclusion, this study describes a lethal 212kb deletion with pleiotropic effects on two different genes, one resulting in fetal death in homozygous state (BMPER), and the other increasing growth (BBS9) in heterozygous state. We provide strong evidence for balancing selection resulting in an unexpected high frequency of a lethal allele in the population. This study shows that the large amounts of genomic and phenotypic data routinely generated in modern commercial breeding programs deliver a powerful tool to monitor and control lethal alleles much more efficiently.

Author summary

We report a large deletion within the *BBS9* gene that induces late fetal mortality in homozygous affected animals in a commercial pig population. This late fetal mortality causes the fetus to become encapsulated and desiccated during the remaining time of the pregnancy, a process called mummification. The unusually high carrier frequency for this lethal deletion (10.8%) likely results from its strong positive association with growth rate in heterozygous individuals, an important selection trait in the pig breeding industry. Interestingly, we show that the positive effect on growth is induced by a heterozygous loss-of-function of the *BBS9* gene,

associated with obesity in human and mouse. However, late fetal mortality is induced by insufficient expression of the *BMPER* gene located directly downstream of the deletion which affects its regulatory elements required for gene expression. Together, our study shows an unique example of allelic pleiotropy in which one allele (deletion) is responsible for both increased growth and late fetal mortality by affecting two different genes.

5.1 Introduction

Domesticated animals are excellent models to study the effect of inbreeding on fitness, and the role of selection in inbreeding depression. Breeding of domesticated animals increases inbreeding by applying artificial insemination that allows breeding populations to be sired by a small number of elite males. The frequency of deleterious alleles including recessive lethal alleles can rise in populations as a consequence of drift due to small effective population size, but also due to selection (Leroy 2014). Inherited defects usually derive from unique “founder” mutations (Yin et al. 2014). Especially in cattle breeds, several high frequency lethal alleles have been described (Sahana et al. 2016; Hoff et al. 2017) reaching carrier frequencies up to 32% (Kadri et al. 2014), that can be traced back to prime bulls that were used extensively in the past decades. However, the effect of individual sires on the population depends on the breeding goal and the structure of the breeding program. In cattle breeding, the genetic contribution of a single bull can be extreme, producing up to hundreds of thousands of daughters. In pig breeding, however, drift effects are expected to be less severe because recessive lethal alleles from founder boars are less likely to rise in frequency very rapidly, because of a lower male selection intensity compared to cattle breeding (Knol et al. 2016).

The role of random drift and/or selection in increasing the frequency of deleterious variants is complex. When effective population size is small, drift effects can result in less effective selection (Bosse et al. 2015). Interestingly, the number of lethal variants found at relatively high frequency in commercial pig populations appears to be low (Haggman and Uimari 2016; Derks et al. 2017; Howard et al. 2017a). The relative paucity of high-frequency deleterious alleles in pig and chicken, species that generally show a more gender-balanced selection (Bouquet and Juga 2013; Knol et al. 2016), and larger effective population size compared to cattle breeds (Hidalgo et al. 2016; Kelleher et al. 2017), raises the question why still some alleles rise to moderate frequency despite having a very clear adverse effect. Heterozygote advantage for traits selected in commercial populations provides a tantalizing alternative hypothesis (Hedrick 2015). In cattle, various instances of balancing selection have been described, driving deleterious alleles to higher population frequencies (Fasquelle et al. 2009; Kadri et al. 2014). In pigs, similar observations were made involving a transposable element (L1) insertion with positive effect on litter size, but negative consequences for boar fertility (Sironen et al. 2012).

In a previous study we identified various recessive lethal alleles in three pig breeds (Derks et al. 2017), but the majority of these lethal alleles were found at low frequencies. One recessive lethal haplotype, however, was found at moderate frequency (~9% carrier frequency) causing a significant increase in foetal mortality at mid- to late-gestation and resulting in a high fraction of mummified piglets in a Large White commercial population. The strong deleterious nature of the allele and the high frequency suggests a factor other than drift driving this haplotype to high frequency.

In this study, we report evidence of balancing selection on a recessive lethal 212kb deletion within the *BBS9* gene with antagonistic effects on fertility and growth. The allele affects fertility by causing early fetal death in homozygous progeny, resulting in mummified piglets. The same allele increases growth rate and feed intake for carrier animals compared to non-carrier animals. We propose that the deletion is maintained at moderate frequency in the Large White breed because of its association with this positive effect, despite it being lethal in homozygous state.

5.2 Results

5.2.1 A haplotype inducing foetal lethality segregates at moderate frequency in a Large White pig population

Genomic loci that harbour recessive lethal alleles can be identified by searching for haplotypes showing reduced or missing homozygosity. In this study, we analysed a previously identified recessive lethal haplotype on pig chromosome 18 (SSC18: 39.25–40.1 Mb) using 23,722 Large White animals from a single purebred sow line genotyped on the Porcine50K SNPchip (Sscrofa11.1 build). The haplotype frequency is estimated at 5.4% (10.8% carrier frequency, Table 5.1), showing that the haplotype is segregating at moderate frequency in this Large White population. In total, we expect 55 homozygote carriers for the SSC18 haplotype within the population. However, no homozygous del/del animals were observed, supporting that all copies of the haplotype carry the recessive lethal variant exhibiting complete penetrance for homozygous animals. We also observe a significant reduction in total number born (19.5%) and liveborn individuals (19.3%) for carrier-by-carrier matings (CxC) compared to carrier-by-non-carrier matings (CxNC). Moreover, we found an approximate fivefold increase in mummified piglets (Table 5.1). The difference between stillborn and mummies lies in the moment the foetus dies: The term 'mummy' is used for a foetus that dies mid-to-late-gestation (e.g. second to third trimester) and is subsequently encapsulated and desiccated during the remainder time of the pregnancy. A foetus that dies near the end of gestation or perinatally is identified as 'stillborn'. The reduction in total number born is slightly lower than the expected 25% based on the 1:2:1 genotype distribution expected from CxC matings. About 73% of the CxC progeny is heterozygous for the SSC18 haplotype, corresponding to the 1:2 genotype ratio expected for CxC matings that lack homozygous offspring, significantly different compared to the normal 1:2:1 Mendelian ratio (p = <0.00001). Based on the carrier frequency, we estimate that about 1.17% of the litters within this breed are affected by the SSC18 haplotype, producing affected animals ('mummies'), and resulting in reduced litter sizes (on average 3.08 piglets per CxC litter).

Table 5.1 SSC18 haplotype characteristics and phenotypic effects. Difference is the percent difference in the average total number born (TNB), number born alive (NBA), and mummified piglets (MUM) for C x C (carrier-by-carrier) and C x NC (carrier-by-non-carrier) matings.

Position, Mb	SSC18: 39.2–40.1
Number of markers	25
Starting marker	ASGA0079708
Ending marker	ALGA0098146
Homozygotes expected (trio)	55
Homozygotes observed	0
Exact binomial test	1.12e-27
Haplotype frequency %	5.42
Carrier frequency %	10.84
C x C matings	154
Genotyped C x C progeny	218
Heterozygote C x C progeny	159 (72.9%)
Avg. TNB (difference %)	12.86 (-19.5%)
Avg. NBA (difference %)	11.69 (-19.3%)
Avg. MUM (difference %)	1.62 (476.4%)
Genes in window	BMPER, BBS9

5.2.2 Genotyping the offspring of carrier-by-carrier matings confirms early lethality of homozygous animals

We tracked five recent CxC matings. Four pregnancies reached full term, while one resulted in spontaneous early abortion of the entire litter (Table 5.2). The four full-term litters produced 49 liveborn, 7 stillborn, and 14 mummified piglets. Each of these four litters produced at least 2 mummified piglets (maximum 5), significantly more than what is normally observed in this breed (on average 0.35 mummified piglets per litter, *p* = 0.0027). Among the total of 48 genotyped liveborn and stillborn siblings (8 siblings were not genotyped), 16 were non-carriers, 30 were heterozygous (62.5%), and two were homozygous for the SSC18 haplotype, close to the expected 1:2 genotype ratio caused by missing homozygous offspring (S1and S2 Tables). Among the two "fresh born" homozygous animals (i.e. piglets surviving at least until around birth), one was a stillborn piglet, the other was a liveborn but very weak piglet, that died shortly after birth.

Table 5.2 Tracked CxC matings for the SSC18 haplotype. Phenotypes and genotypes of 4 litters from CxC matings from two different farms. The number of successfully genotyped individuals are indicated between parentheses for each birth type. Litter CC3 contains two fresh born homozygous individuals. An overview presenting the haplotypes and carrier status of the four litters is provided in S3 and S4 Tables.

Litter	Farm	Parity	Liveborn	Stillborn	Mummified	# Non-carriers	# Carriers	# Confirmed homozygotes
CC1	1	5	10 (6)	1 (1)	4 (0)	2	5	-
CC2	2	1	12 (11)	0	3 (1)	3	8	1
CC3	1	5	17 (15)	3 (3)	2 (0)	7	9	2
CC4	2	3	10 (10)	3 (2)	5 (1)	4	8	1

We confirmed the homozygous status for two mummified piglets with sufficient call rate (call rate > 0.8, S1 Table), the other mummified piglets yielded insufficient DNA quality to perform genotyping and phasing (call rate < 0.8, S1 and S2 Tables). Next, we collected eight mummified piglets from one farm for phenotypic evaluation (including X-rays, S1 Fig), the other six mummified piglets were measured (length), but not stored. The approximate age when a mummified pig has died can be determined based on the length (crown to rump) and weight. The majority of the mummified piglets die approximately in the second half of the second trimester of pregnancy (50–70 days), based on the length (100–200 mm) and weight (100–190 gram) of the mummified piglets (S1 and S2 Tables, S1 Fig). Three mummified piglets from one litter (litter ID: CC4) died later in gestation as was evident from a larger size and weight (S1 Table). However, we cannot confirm the homozygous status for the SSC18 haplotype, since these animals could not be successfully genotyped due to poor DNA quality. Together these results support a broad range in the time of death between homozygous animals (supporting variation in penetrance), ranging from 50 days in gestation to 24 hours post-partum.

5.2.3 Carriers exhibit a 212kb deletion affecting the *BBS9* gene

To identify candidate causal mutations, we analysed whole genome sequence data from 73 individuals from the same Large White population and identified 10 carrier animals for the SSC18 haplotype (S5 Table). We first annotated loss-of-function and (deleterious) missense mutations within and surrounding the haplotype region (+/- 5 Mb) uniquely found in the SSC18 haplotype carriers. However, none of the mutations were predicted to have high impact (Variant Effect Predictor, build 90 (McLaren et al. 2016)). Next, we assessed the presence of structural variation within the same region and identified a large deletion in complete LD with the SSC18 haplotype of approximately 212kb (position 39,817,373 to 40,029,300), spanning a part of the *BBS9* gene (Figure 5.1A and 5.1B). The deletion is supported by both split-reads and discordantly mapped pairs in carrier samples (S2 Fig). Moreover, carrier animals show reduced signal intensities (referred to as Log R Ratio; Fig 5.1A, S3 Fig), and increased homozygosity for four markers on the Porcine50K SNPchip located within deletion, caused by the absence of a second haplotype for the deletion region. In addition, several markers neighbouring the deletion show an excess of heterozygosity, caused by the absence of homozygous del/del animals.

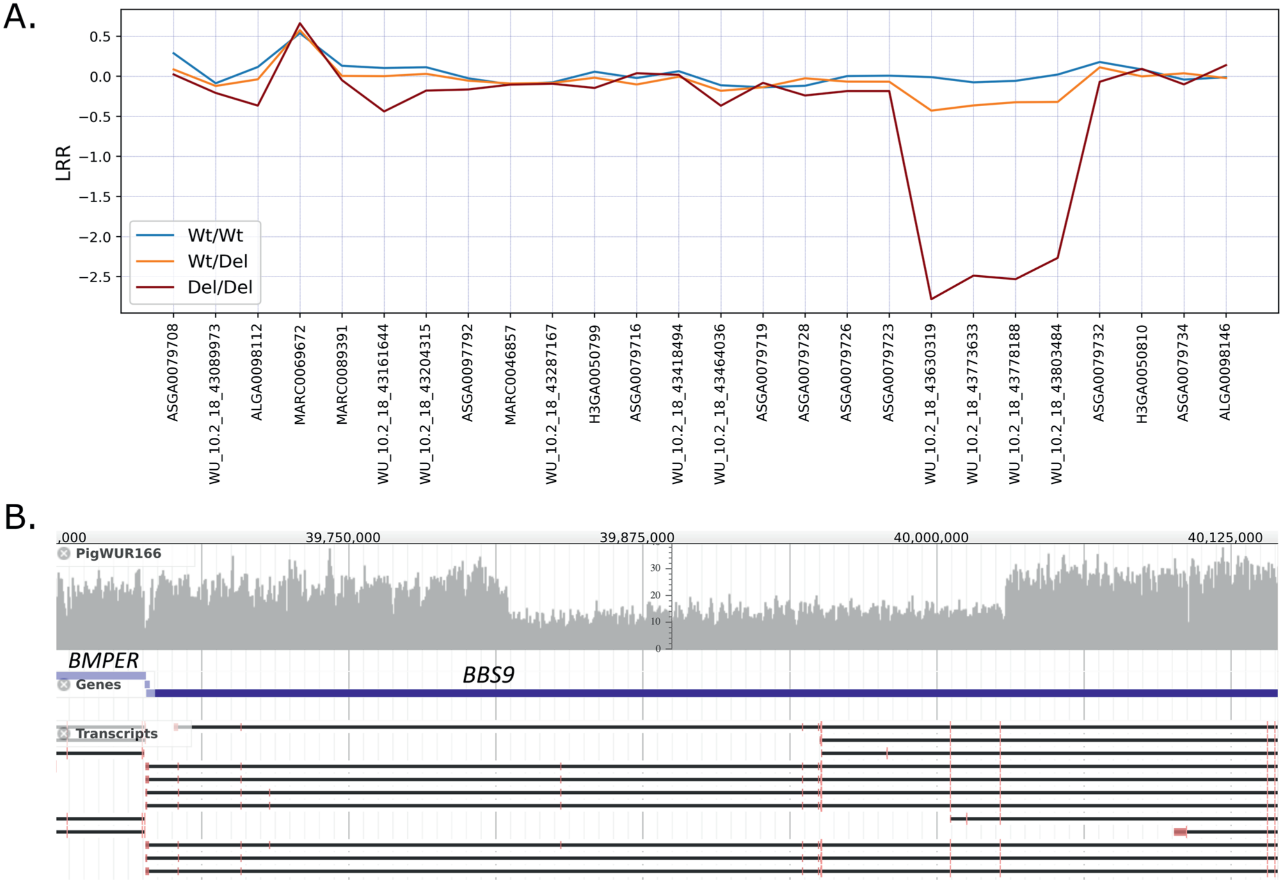
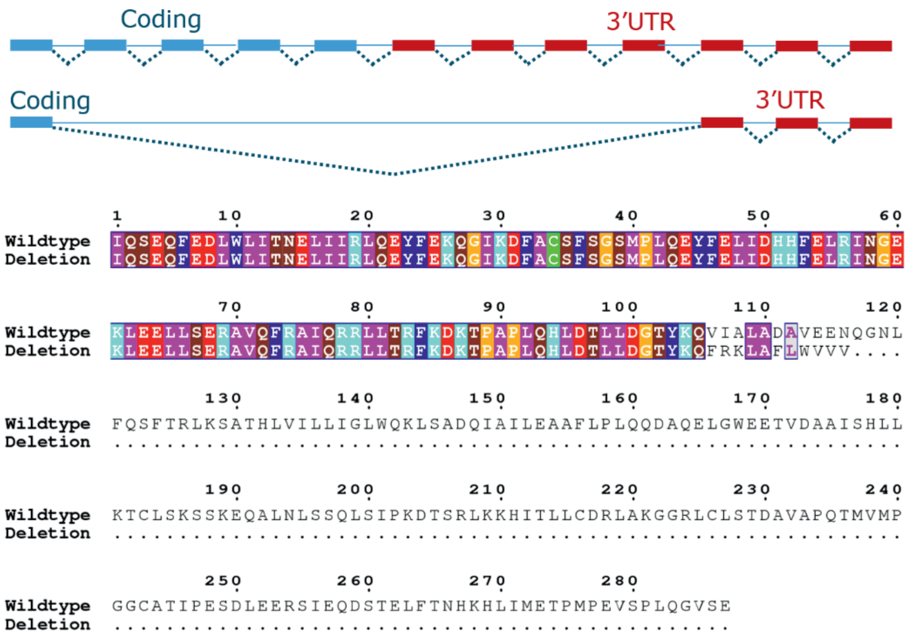


Figure 5.1 A) Log R Ratio (LRR) signal intensities on the 50K SNPchip for homozygotes (del/del) carriers (wt/del), and non-carriers (wt/wt). Four markers within the 212kb deletion show reduced LRR intensities. B) Screen capture of the alignment of carrier animal PigWUR166. The aligned region on SSC18 shows reduced coverage in the deletion within the *BBS9* gene.

5.2.4 SSC18 deletion produces a truncated BBS9 protein

We analysed RNA-seq data from one carrier animal in eight different tissue types (sample: PigWur166, S6 Table) to investigate the impact of the deletion on the expression of *BBS9*. Moderate gene expression levels for *BBS9* were observed for the majority of the examined tissues, except for muscle, and with highest gene expression in testis (S6 Table). We evaluated the effect of the deletion on the *BBS9* mRNA and show that the deletion induces skipping of 4 coding, and 4 3'UTR exons for the *BBS9* canonical transcript (Figure 5.2, RefSeq ID: XM_021079336.1), resulting in direct splicing from exon 19 to exon 28 (3'UTR). The mutant transcript results in a frameshift introducing 11 novel amino acids before a premature stop codon, generating a truncated *BBS9* protein of 694 amino acids (including 11 novel amino acids) instead of the wild type 865 amino acids. This truncated BBS9 protein will likely be non-functional (Figure 5.2), supported by pathogenic mutations identified in humans affecting the same C-terminal tail of the BBS9 protein (Nishimura et al. 2005; Shaheen et al. 2016). Moreover, the affected protein coding exons exhibit a negative subRVIS score, indicating intolerance to loss-of-function mutations (Gussow et al. 2016). Finally, we evaluated the expression of *BBS9* using a RT-qPCR on 8 carrier and 10 non-carrier samples from whole blood using primers that target exons located within the deletion. The results show a 50% lower expression of the wild-type *BBS9* gene in carrier animals (S4 Fig).



```
           1        10        20        30        40        50        60
Wildtype  IQSEQFEDLWLITNELIIRLQEYFEKQGIKDFACSFSGSMPLQEYFELIDHHFELRINGE
Deletion  IQSEQFEDLWLITNELIIRLQEYFEKQGIKDFACSFSGSMPLQEYFELIDHHFELRINGE

                  70        80        90       100       110       120
Wildtype  KLEELLSERAVQFRAIQRRLLTRFKDKTPAPLQHLDTLLDGTYKQVIALADAVEENQGNL
Deletion  KLEELLSERAVQFRAIQRRLLTRFKDKTPAPLQHLDTLLDGTYKQFRKLAFLWVVV....

                 130       140       150       160       170       180
Wildtype  FQSFTRLKSATHLVILLIGLWQKLSADQIAILEAAFLPLQQDAQELGWEETVDAAISHLL
Deletion  ............................................................

                 190       200       210       220       230       240
Wildtype  KTCLSKSSKEQALNLSSQLSIPKDTSRLKKHITLLCDRLAKGGRLCLSTDAVAPQTMVMP
Deletion  ............................................................

                 250       260       270       280
Wildtype  GGCATIPESDLEERSIEQDSTELFTNHKHLIMETPMPEVSPLQGVSE
Deletion  ...............................................
```

Figure 5.2 BBS9 "Wild-type" (top) and mutant (bottom) transcripts. The deletion transcript skips four coding and four 3'UTR exons, resulting in a frameshift (indicated with an arrow in the alignment) introducing 11 AAs before a preliminary stop codon.

5.2.5 SSC18 deletion lowers *BMPER* expression by affecting cis-regulatory elements

To evaluate the impact of the deletion on the downstream *BMPER* gene we investigated possible allelic imbalance for the *BMPER* gene within the same carrier animal. The *BMPER* gene is highly expressed in lung, while moderately expressed in the other tissue types (S6 Table). One heterozygous coding synonymous mutation within the fourth exon of the *BMPER* canonical transcript (XM_013990842.2) was used to test for allelic imbalance. Interestingly, we observed a three-fold higher expression of the *BMPER* allele for the wild-type haplotype (T allele) compared to the del haplotype (in lung tissue, Table 5.3). By contrast, three homozygous wild-type animals showed no allele specific differences in expression for the *BMPER* gene (S7 Table), suggesting that the region affected by the 212kb deletion contains *BMPER* cis-regulatory elements. To support the presence of *BMPER* regulatory elements within the deletion we aligned liver ChipSeq (H3K27Ac, H3K4Me3) data (Villar et al. 2015) to the Sscrofa11.1 genome build. Two strong enhancer peaks are observed within the deletion region, while only weak signals are observed outside the deletion region (S5 Fig). In addition, the sequence of the 212kb deletion was mapped to the human genome to identify the homologous sequence on the human genome (GRCh38: Chr7:33.50–33.71). This region contains several conserved regulatory elements, identified from the Regulatory Element Database (Sheffield et al. 2013), one non-coding RNA (LOC105375227), and several enhancer sites, of which at least two are annotated to enhance *BMPER* expression according to the human EnhancerAtlas (Gao et al. 2016).

Table 5.3 Allele specific expression of the *BMPER* gene for a SSC18 carrier animal. One heterozygous coding synonymous SNP within the fourth exon of the BMPER canonical transcript (XM_013990842.2) was used to test for allelic imbalance.

Locus	Gene	Del-allele	Alt-allele	Del-Count	Wt-Count	Ratio	FDR-p
18:39594479	BMPER	C	T	24	73	0.753	2.35e^{-05}

5.2.6 Tracing the origin of the deletion

To investigate the origin of the deletion, we analysed the frequency of the deletion over the last decades. The first born animals within our genotyped set are from

February 2006, allowing the tracking of the frequency of the deletion over the past decade. The number of genotyped animals was lower in the period 2006–2010. However, we genotyped over 320 animals in the (live) population from 2008 onwards, providing reliable frequency estimates (S8 Table). The SSC18 haplotype carrier frequency was high (>15%) over the period 2006–2010 (maximum 20% in 2008) and then decreased to a relative stable ~10% carrier frequency from 2012 onwards (Figure 5.3).

The Large White population under study has been created out of the consolidation of a number of Dutch breeding organizations around the turn of the last century (Hoving et al. 2017). During the consolidation phase, which resulted in merging of populations and phasing out of other populations, sperm of breeding boars was deposited at the Dutch Centre for Genetic Resources (CGN). The current Large White pure line descends from two different populations, the StamBoek-Z and the Dumeco-W line (Hoving et al. 2017). Both breeds were merged around 2003 to form the current Large White breeding line. From the 11 StamBoek-Z boars available at CGN, none were carrier of the deletion. However, from the 56 genotyped Dumeco-W boars available at CGN, five were carrier for the deletion haplotype (8.9%). These boars were born in 2000 and 2001 (S9 Table), showing that the deletion derives from this ancestral line and has been maintained in the Large White population for the past eighteen years (~ 15 generations).

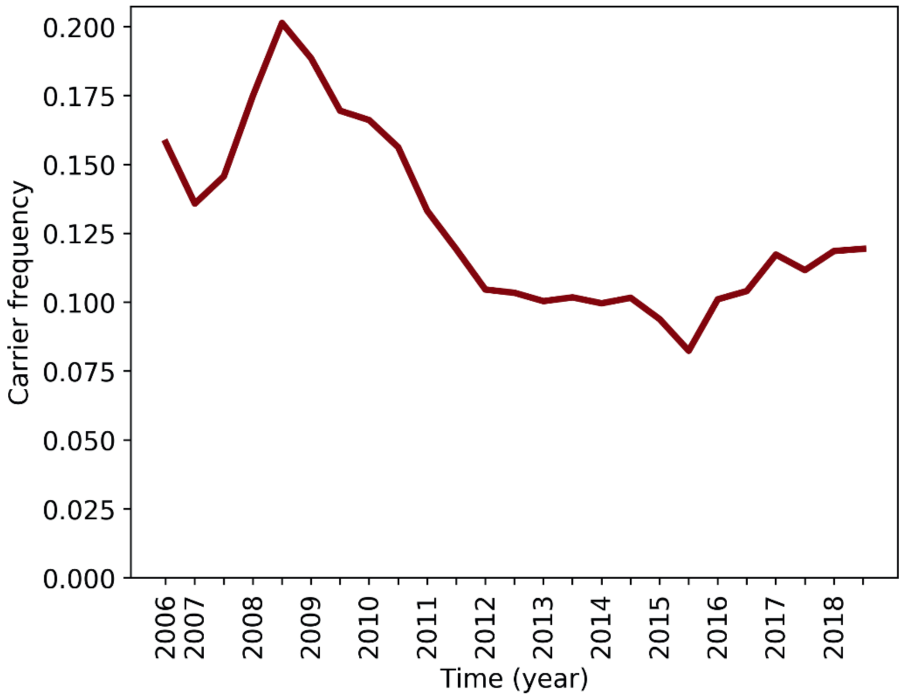


Figure 5.3 SSC18 carrier frequency from 2006–2018. The frequency has changed significantly over the past 12 years (p = 0.012).

5.2.7 Carriers of the deletion have increased growth rate and feed intake

We examined whether the current carrier frequency is purely the result of genetic drift, or whether carriers exhibit selective advantage for important traits within the breeding program. We first simulated genetic drift for a lethal recessive allele in the current Large White population (S7 Fig). The results show that lethal alleles can reach allele frequencies up to about 10% by drift alone (although extremely rare), at which the lack of homozygotes is preventing further increase. Next, we tested whether deletion carriers exhibit heterozygote advantage, by performing an association study for both carrier and non-carrier animals using deregressed estimated breeding values (DEBVs) for 16 production traits available from the Topigs-Norsvin breeding program (Table 5.4)

Table 5.4 Traits significantly associated with heterozygous carriers of the deletion. Effect shows the direction of the association, se shows the standard error. Table shows increased DEBVs for growth rate (TGR: growth rate in test period, ~25-120Kg, LGR: lifetime growth rate),

daily feed intake (DFI), and litter mortality (LMO), while decreased DEBV for litter birth weight (LBW, grams), loin depth (LDE), and longevity (LGY) are observed for carriers. The symbols "+" and "-" indicate positive and negative effects. The effect on DFI can be considered both positive and negative. If TGR is increasing, DFI tends to increase a bit. However, it should not increase too much because it will affect feed conversion. An overview of all traits tested is provided in S10 Table.

Trait (unit)	Non-carriers	Carriers	P	-log10(P)	effect	se
TGR (gr/day) $^{+}$	15013	1605	0.000046	4.34	11.46	2.81
LDE (mm) $^{-}$	15011	1598	0.000198	3.70	-0.45	0.12
LGR (gr/day) $^{+}$	15116	1616	0.000315	3.50	6.40	1.77
LBW (gram) $^{-}$	6945	824	0.001232	2.91	-16.67	5.16
LMO (%) $^{-}$	7345	871	0.001248	2.90	0.67	0.21
DFI (gr/day) $^{+/-}$	14671	1567	0.006764	2.17	30.61	11.30
LGY (parity) $^{-}$	7250	856	0.024828	1.61	-0.08	0.04

The carriers grow faster (TGR and LGR), have smaller loin depth (LDE), produce litters that are lighter (LBW), show higher mortality in their litters (LMO) and have a higher feed intake (DFI) when compared to the non-carriers. Selection on growth has not significantly changed in the last decade, and there is consistent increase in genetic progress for growth in this time period (S8 Fig). To further support a balancing scenario, we evaluated the difference in the total selection index (TSI) between the carrier and non-carrier group for all animals born in 2017. Animals are ranked based on this selection index to select the top animals to produce the next generation. We observe a 2.7% higher TSI (on average) for carriers compared to non-carriers (S11 Table), caused by the positive effect on growth, that outweighs the negative effect on other traits. Next, we simulated the long term effect on the SSC18 carrier frequency based on the current heterozygous advantage and frequency (Figure 5.4, S9 Fig). We observe a decrease in carrier frequency in the first generations due to the loss of homozygotes, which outweighs the heterozygous advantage perceived in the selection index. However, at approximately 6% carrier frequency, the heterozygous advantage compensated for the loss of homozygous offspring, reaching a trade-off at this point. Moreover, carriers show 12.4% higher breeding values for growth compared to non-carriers (S11 Table), and we show that the carrier frequency can rise up to 22% if selection would be exclusively on growth (S10 Fig). Together these results support a balancing selection scenario showing heterozygote advantage for growth rate (Figure 5.5), an important selection trait in the pig breeding industry.

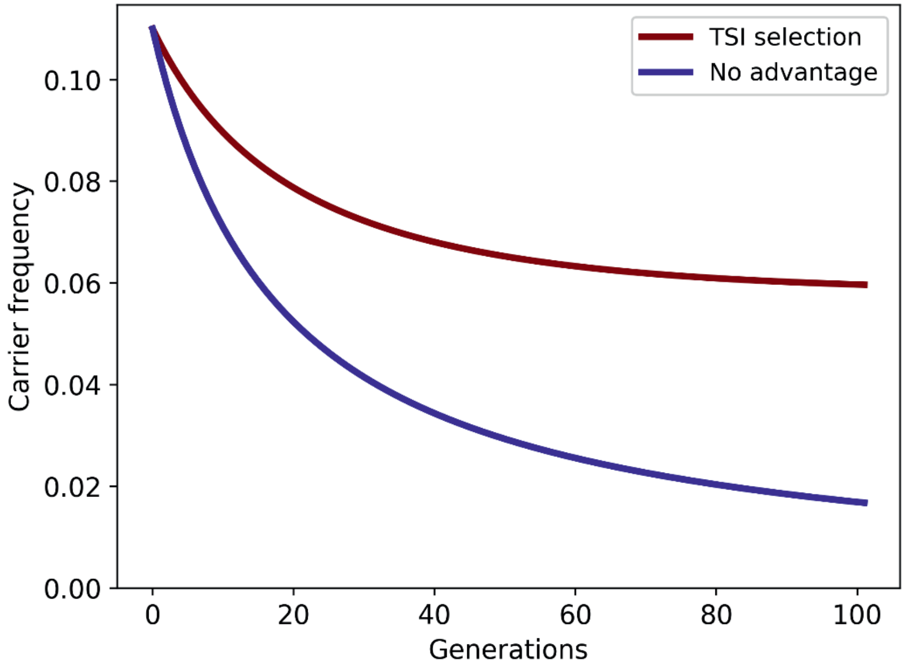


Figure 5.4 Simulation of the SSC18 carrier frequency with current selective advantage over 100 generations starting with current carrier frequency (11%). Figure shows a decrease in carrier frequency in the first generations due to the loss of homozygotes which outweighs the heterozygous advantage (~3%) perceived in the selection index (TSI). Figure shows a trade-off at approximately 6% carrier frequency at which the heterozygous advantage is compensated by the loss of homozygous offspring.

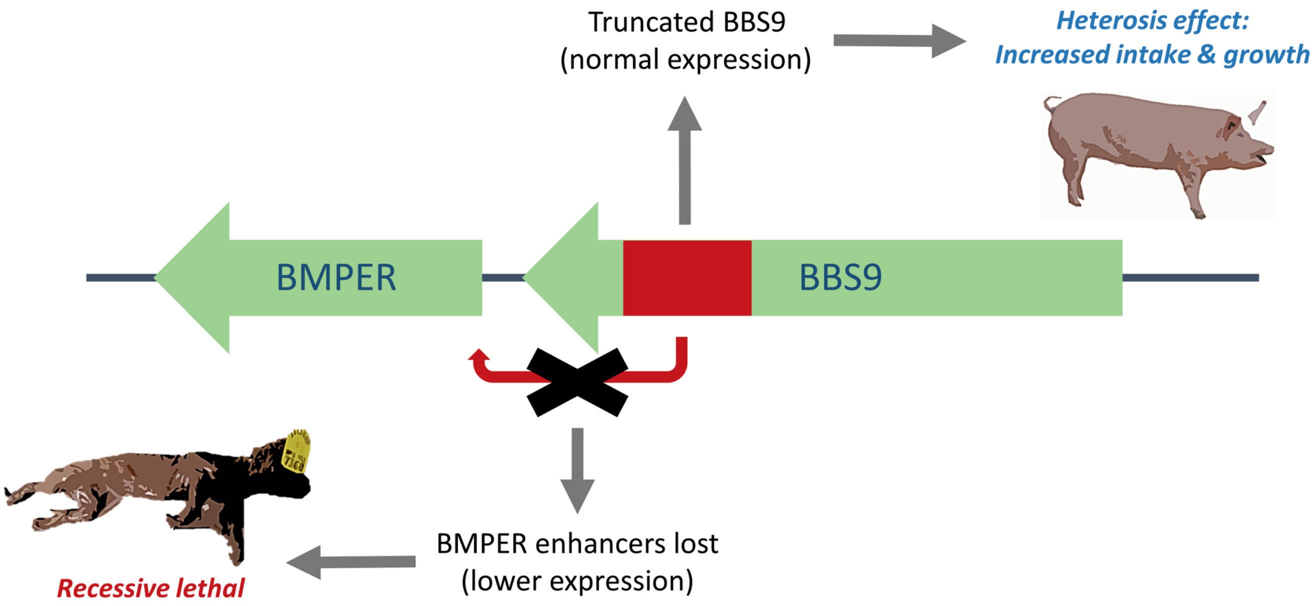


Figure 5.5 Schematic representation of the SSC18 deletion affecting *BMPER* gene expression and *BBS9* protein structure. A heterozygous loss of function of the *BBS9* gene results in increased growth rates, while reduced expression of the *BMPER* gene results in foetal mortality in homozygous del/del animals

5.3 Discussion

Livestock populations with small effective population size can lead to the spread of recessive lethal alleles, in which the effects of drift tend to dominate over the effect on selection (Glemin 2003). On the other hand, purging lethal alleles is more efficient in populations with small effective population size, which seems to contradict the relative lack of purging for the current allele under study over such long time period (almost two decades). One possible-partial-explanation is that fertility, a complex trait, is influenced by many genetic and environmental factors. As a consequence, fertility traits generally have low heritability's (Zak et al. 2017). One of the important determinants of fertility is prenatal mortality (van der Lende et al. 2001). However, prenatal mortality caused by recessive genetic defects is often difficult to capture within the pig breeding values (smaller litters will only be generated if two carriers mate), especially if the lethal variant is segregating at relative low frequency in the population.

We report a 212kb deletion that causes death of homozygous fetuses, and which also shows a positive effect on growth rate and feed intake in heterozygous pigs (Figure 5.5). We show that the allele has been segregating in the population for at least 18 years, despite its detrimental lethal effects in its homozygous state. The frequency of the deleterious allele was higher a decade ago, likely caused by genetic drift, because no significant changes in selection pressure for production traits have been applied during this time period. The balancing nature of the allele is clear from the higher TSI index of carrier animals. The 212kbp deletion clearly shows a net positive effect on the chances of pigs becoming selected in the breeding program, since the TSI is used to rank selection candidates. Although lower fertility is not captured in the TSI or even breeding values related to fertility directly, it is clear that the lower number of piglets born from CxC matings will result in a decreased fitness of these crosses. As a consequence, a balance is expected to arise between positive and negative selection, and indeed, the heterozygous advantage captured in the selection index compensates for the loss of homozygotes at approximately 6% carrier frequency. This frequency is somewhat lower than the current and past observed frequency. One partial explanation for this discrepancy is that we overestimate the number of CxC matings in our simulations (assuming random matings), because in practice matings between related individuals are avoided. Furthermore, during the 20th century, selection was mainly applied on growth and carcass traits (Merks 2000), indicating that the heterozygote advantage for carriers was likely stronger in the past. Interestingly, the balance between positive and negative selection would explain a rapid increase in the population under strong selection for growth traits, as has been the case for pig breeding lines at least since the introduction of modern breeding techniques.

This increase in growth rate (with the most significant effect in the test period from 25–120 Kg) and feed intake, likely results from a heterozygous loss-of-function of the *BBS9* gene. The BBS9 protein is part of the BBSome complex, and is required for ciliogenesis (Novas et al. 2015). BBS9 is the central organizational component of the BBSome, having direct interactions with BBS1, 2, 5 and 8 (Klink et al. 2017). Loss-of-function mutations in human *BBS9* and other members of the BBSome cause Bardet-Biedl syndrome, associated with a series of clinical features including obesity, renal anomalies, and retinopathy, with the obese phenotype as one of the key features of Bardet-Biedl syndrome patients. Studies have hypothesized that cilia defects are likely to affect feeding and satiety, causing an increased appetite and lack of satiation (Novas et al. 2015). In addition, heterozygous carriers of a mutant BBS allele in humans show increased levels of obesity, without showing any of the other Bardet-Biedl syndrome features (Croft et al. 1995), analogous to the observed phenotype in carriers of the *BBS9* deletion. Mouse null-mutants in genes that form the BBSome complex have been associated with similar phenotypic features including obesity, lower birth weights, and partial embryonic lethality (Kulaga et al. 2004; Funari et al. 2010), again supporting the *BBS9* role in increased growth rate, and the lower birth weight. We cannot completely exclude, however, that other genomic factors, in high LD with the 212kb deletion, contribute to the observed phenotype as well.

The question remains which gene or regulatory element is causal for the early death of homozygous individuals. We expect that the deletion (in homozygous state) leads to a complete loss-of-function of the *BBS9* gene, and decreased expression of the *BMPER* gene by affecting *BMPER* enhancer elements. Enhancers are important drivers of transcription and loss of enhancer elements can lead to decreased expression of the associated gene. Naturally occurring knock-outs of the *BBS9* gene does not result in fetal lethality in human (Nishimura et al. 2005) and is therefore not likely to be causal for the lethal phenotype. Instead, the downstream *BMPER* gene is a much stronger candidate since *BMPER* null mutants result in prenatal lethality with skeletal malformations in both mice and human (Ikeya et al. 2006; Funari et al. 2010), marking the *BMPER* gene as the likely candidate underlying this phenomenon. We hypothesize that the deletion affects *BMPER* enhancer elements, resulting in insufficient expression of the *BMPER* gene in homozygous state. The lower expression of the BMPER gene is supported by allele specific expression for the non-deletion haplotype in carriers, not observed in individuals only carrying wild-type haplotypes. However, other BBS proteins, part of the BBSome, do cause (partial) embryonic lethality (i.e. *BBS4* (Kulaga et al. 2004) and *BBS7* (Zhang et al. 2012)), we can therefore not exclude the possibility that a complete loss of a functional BBS9 protein contributes to the early lethality as well. Together these results support that the deletion affects *BMPER* regulatory elements resulting in allelic imbalance for the *BMPER* gene in carrier animals. Therefore, severe downregulation of the *BMPER* gene is expected for del/del animals causing fetal death.

This work describes a striking example of balancing selection in pigs, maintaining a recessive lethal allele that shows pleiotropic effects on fertility and growth traits at moderate frequency in the population. Other examples in pigs include the Porcine Stress and Pale Soft Exudative Meat Syndrome, caused by a homozygous missense mutation in the *RYR1* gene, while heterozygotes show increased muscle mass (Fujii et al. 1991). A second example is a LINE insertion in the *SPEF2* gene causing increased litter size in sows but decreased fertility in boars (Sironen et al. 2012). Moreover, several instances of balancing selection have been described in domestic cattle breeds among which a 660kb deletion causing embryonic lethality in homozygotes, while having increased milk yield in heterozygotes (Fasquelle et al. 2009; Kadri et al. 2014). Identifying balancing selection on lethal alleles can be challenging, as the only consequence observed is a (somewhat) lower fertility in the parental animals, lacking affected (liveborn) individuals. We expect that this type of balancing selection might be more prevalent within pig populations than previously thought, especially for the somewhat higher frequency lethal alleles, which are less likely to be purely the result of drift effects. Moreover, the relatively subtle effects found in this study could only be made apparent because phenotypic data derived from a very large number of pigs was available.

5.4 Conclusions

In this study we report a 212 kb deletion with antagonistic effects on fertility and growth. We show that homozygotes for the deletion die around mid- to late-gestation, becoming mummified. Compared to other lethal alleles identified in this population, the deletion seems to be maintained at moderate frequency (10.8%) in the population. This moderate carrier frequency is likely not a result of random drift effects, as heterozygotes for the deletion-haplotype show, despite a lower birth weight, increased growth rate, and feed intake, important traits in the breeding goal. The balancing scenario observed, most likely, is a consequence of pleiotropic effects of the deletion on two different genes affecting fertility (*BMPER*) and growth (*BBS9*). The large amount of genotype data accumulating in modern breeding schemes applying genomic selection in combination with the large amount of phenotypic data deliver a powerful tool to monitor and control deleterious alleles much more efficiently.

5.4 Methods

5.4.1 Ethics statement

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

5.4.2 Animals, genotypes and pre-processing

The dataset consists of 23,722 purebred Large White animals. The animals were genotyped on the Illumina GeneSeek custom 50K SNP chip (Lincoln, NE, USA). Animals with a frequency of missing genotypes > 0.20 were removed. We discard markers that did not meet following filtering criteria: A minimum call rate of 0.85, a minor allele frequency > 0.01, and a Hardy-Weinberg proportions exact test p-value below $P < 10^{-6}$. Moreover, markers with unknown location on the Sscrofa11.1 genome build were discarded, leaving 42,288 markers after filtering. All steps were performed in Plink v1.90b3.30 (Purcell et al. 2007).

5.4.3 Haplotype phasing

We performed haplotype phasing and imputation of missing sites in Beagle4.1 with parameter for effective population size set to 195, other settings were default (Browning and Browning 2007). Reference and test phased VCF files were merged using bcftools 1.3–27-gf31e888 (Li et al. 2009).

5.4.4 Identification of missing homozygote haplotypes

We tested the SS18 haplotype for the expected number of homozygotes using both parents haplotype information (sire, and dam) with the formula described in Fritz et al., 2013 (Fritz et al. 2013). An exact binomial test was applied to test the number of observed homozygotes with the number of expected homozygotes. The haplotype was considered significantly depleted if $P < 5 \times 10^{-3}$. The difference in Mendelian ratios for CxC compared to CxNC matings was tested using a Chi-Square test.

5.4.5 Pseudo genotyping for SSC18 deletion

To genotype animals directly for the SSC18 deletion, we first calculated LRR normalized signal intensities using PennCNV analysis software (Wang et al. 2007). We built a classifier with 5 features: the LRR signal intensities for the four overlapping markers within the deletion (WU_10.2_18_43630319, WU_10.2_18_43773633, WU_10.2_18_43778188, WU_10.2_18_43803484), and the average LRR signal intensity over these four markers. Next, we applied logistic regression to distinguish carrier from non-carrier animals using the sci-kit learn Python library (Pedregosa et al. 2011) (S6 Fig).

5.4.6 Phenotypic effects associated with lethal haplotypes

We examined the SSC18 haplotype for records on TNB, NSB, and MUM listed for all C x C, and C x NC matings identified in the phenotypic records, the order of C x NC matings does not reflect the sex of the parent animal and is both carrier boar and carrier sow combined. We used a Welch's t-test to assess whether the phenotypes from the C x C matings differ significantly from C x NC matings. A p-value < 0.05 was considered significant.

5.4.7 WGS analysis and candidate variant identification

The dataset consists of 73 whole genome sequenced Large White individuals with a total volume of 1.77 Tbp (tera base pairs) from 15.539 billion paired-end reads, ranging from 100–150 bp in length (S5 Table). The data was sequenced on Illumina Hiseq 2000. We used sickle software for quality trimming of raw reads. Next we aligned the sequences to the Sscrofa11.1 genome build using BWA-MEM version 0.7.15 (Li and Durbin 2009) with an average mappability of 96.11% and a sample coverage ranging from 6.6–22.7X (10X average). Samtools dedup function was used to remove PCR duplicates (Li et al. 2009). GATK IndelRealigner was used to perform local realignments around indels (McKenna et al. 2010b). Variant calling was performed with Freebayes v1.1.0 with following settings:—min-base-quality 10—min-alternate-fraction 0.2—haplotype-length 0—min-alternate-count 2 (Garrison and Marth 2012). Variants with phred quality score < 20, and within 3 bp of an indel were discarded (Li et al. 2009). Variants were annotated using the Ensembl variant effect predictor (VEP, release 90) (McLaren et al. 2016). The impact of missense variants was predicted using SIFT (Kumar et al. 2009). The sequenced population was phased using Beagle4.1 (Browning and Browning 2007).

5.4.8 Structural variation analysis

Analysis on structural variation (SV) was performed using Lumpy with default settings (Layer et al. 2014), taking the aligned BAM files as input. Coverage information was calculated for predicted SV events using samtools depth (Li et al. 2009), and added to the VCF format tag using PyVCF. Alignments and SV events were visualized using the JBrowse genome viewer version 1.12.1 (Skinner et al. 2009).

5.4.9 RNA-seq analysis and allele specific expression

We analyzed RNA-seq data on eight different tissues in one SSC18 carrier animal (sample: PigWUR166). In addition, we analyzed two other pigs from Duroc, and Pietrain genetic background on five different tissues. RNA-seq reads were aligned to the Sscrofa11.1 genome build using STAR 2.5.3a (Dobin et al. 2013), generation of transcripts and gene expression levels were achieved with Cufflinks v2.2.1 (Trapnell et al. 2012). We applied the following steps to examine allele specific expression: First, samtools (Li et al. 2009) was used to extract uniquely mapped reads from the BAM alignment files. Next, WASP (van de Geijn et al. 2015) was used to reduce the mapping (reference sequence) bias. Then, GATKASEreadcounter (McKenna et al. 2010b) was used to obtain read counts for reference and alternative alleles at each SNP position. Lastly, a two-sided binomial test with $p = 0.5$ (assuming no bias) and Benjamini-Hochberg false discovery rate (FDR) correction were performed in R v.3.4 at each variant position using the Stats package. The variants with FDR adjusted p-value < 0.05 were considered as allele specific expression variants. Visual examination of the alignments and transcripts was performed in JBrowse (Skinner et al. 2009).

5.4.10 RNA-isolation and RT-qPCR

RNA was extracted from frozen whole blood using the Nucleospin RNA blood kit from Machery Nagel. cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen) with RNA input ~100ng. RT-qPCR was started with: 3.75ul cDNA (1:1), 1.25ul primer forward (2uM), 1.25ul primer reverse (2uM), and 6.25ul MESA blue mix (Eurogentec). RT-qPCR was then performed with a QuantStudio 5 system using the comparative Ct (delta delta Ct) method with GAPDH as housekeeping gene for normalization. Reaction was performed as follows:

1. 50°C 2 min, 1 cycle
2. 95°C 10 min, 1 cycle
3. 95°C 15 sec, 1 cycle
4. 60°C 1 min, 40 cycles
5. 95°C 15 sec, 1 cycle

6. 60°C 1 min, 1 cycle
7. 95°C 15 sec, 1 cycle

Data was analysed with the Quantstudio Design & Analysis Software v.1.4.3. All primers and results are listed in S12 and S13 Tables.

5.4.11 ChipSeq alignment

We downloaded three H3K27Ac, and three H3K4me3 libraries (ArrayExpress accession number: E-MTAB-2633) from liver tissue from three male pig samples described by Villar et al. 2015 (Villar et al. 2015). Data was aligned using BWA-mem (Li and Durbin 2009) and visualized in JBrowse (Skinner et al. 2009).

5.4.12 Frequency over time

We analyzed the frequency of the SSC18 haplotype per half-year starting from 01-jul-2006. We assessed the frequency based on total population (live animals) on each time point by looking at the proportion of carrier and non-carrier animals in the population. The number of animals per time point are provided in S8 Table. We used a One-Way Repeated Measures ANOVA to test whether the frequency differs over time.

5.4.13 Breeding values and association analysis

In this study, we evaluated 16 traits used in the Large White breeding program. Deregressed estimated breeding values (DEBV) were used as a response variable for each trait under study. The estimated breeding value (EBV) was separately deregressed for each trait using the methodology described by Garrick et al (Garrick et al. 2009). The EBV of each animal was obtained from the routine genetic evaluation by Topigs Norsvin using an animal model. The reliabilities per animal for the purpose of deregression were extracted from the genetic evaluation based on the methodology of Tier & Meyer (Tier and Meyer 2004). The heritabilities used for the deregression were also extracted from the routine genetic evaluation. Parent average effects were also removed as part of the deregression process to obtain more accurate estimates of the genetic merit of each individual. Finally, weighting factors based on the estimated reliability of the DEBV were also estimated according to Garrick et al 2009, using a value of 0.5 for the scalar c. To ensure the quality of the DEBV, only animals with a w higher than not equal to zero and a reliability of the DEBV greater than 0.20 were used in the association analyses. The reliability of the DEBV was obtained according to Garrick et al 2009.

Association analyses were performed using the software ASREML (Gilmour et al. 2009) applying the following model:

$$DEBVij\omega = \mu + Ri + aj + eij,$$

where DEBV$_{ij}$ is the observed DEBV for the animal *j*, w is weighting factor for the residual, μ is the overall DEBV mean of the population, R_i is the carrier status of the lethal allele *i*, a_j is the additive genetic effect estimated using a pedigree-based relationship matrix, and e_{ij} the residual error.

5.4.14 Simulating genetic drift

We simulated changes in allele frequency across multiple populations under the model of Wright (Wright 1990). Each allele is associated with a fitness, and we set the fitness to zero for homozygotes (for lethal recessive allele) and fitness to 1 (no negative fitness effect) for carriers and non-carriers. We assume constant population size through time, and matings are simulated randomly at each generation. Changes in allele frequencies are plotted using the R package driftR (https://github.com/cjbattey/driftR).

5.4.15 Balancing selection

Within each generation the top 5% of boars, and top 25% of gilts (based on the TSI selection index value) are used to produce the next generation. We first calculated the average TSI, and estimated breeding values for six important traits in the breeding line (S11 Table). Next, we used the ratio of carrier TSI over non-carrier TSI to estimate the selective advantage in the breeding program. Next, we simulated the long-term allele frequency change (assuming random matings) based on the selective advantage, and the loss of homozygous animals using the Hardy-Weinberg principle. Similar analysis was performed using the selective advantage on growth exclusively.

5.5 Additional files

The online version of this article (https://doi.org/10.1371/journal.pgen.1007661) contains supplementary material, which is available to authorized users.

5.6 Acknowledgements

The authors thank Meenu Bhati for useful input on the allele specific expression analysis. We thank Rita Hoving and Ina Hulsegge for providing the CGN data and information on the consolidation of breeding lines. We thank the FAANG consortium for providing the RNA-seq data. We thank Michel Georges and Carole Charlier for useful comments on this work, Remco Pieters for generating the X-rays, and Richard Crooijmans for library preparation and lab work. We thank Marcel van Kuppenveld, Marc Martens, and Rene Sol for help with sample collection. We thank Egiel Hanenberg for providing EBVs and genetic progress information.

6

Loss of function mutations in essential genes cause embryonic lethality in pigs

Martijn F. L. Derks[1], Arne B. Gjuvsland[2], Mirte Bosse[1], Marcos S. Lopes[3,4], Maren van Son[2], Barbara Harlizius[2], Beatrice F. Tan[1], Hanne Hamland[2], Eli Grindflek[2], Martien A. M. Groenen[1], Hendrik-Jan Megens[1]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. [2]Norsvin SA, Hamar, Norway. [3]Topigs Norsvin Research Center, Beuningen, the Netherlands. [4]Topigs Norsvin, Curitiba, Brazil

PLoS genetics (2019) 15(3): e1008055.

Abstract

Lethal recessive alleles cause pre- or postnatal death in homozygous affected individuals, reducing fertility. Especially in small size domestic and wild populations, those alleles might be exposed by inbreeding, caused by matings between related parents that inherited the same recessive lethal allele from a common ancestor. In this study we report five relatively common (up to 13.4% carrier frequency) recessive lethal haplotypes in two commercial pig populations. The lethal haplotypes have a large effect on carrier-by-carrier matings, decreasing litter sizes by 15.1 to 21.6%. The causal mutations are of different type including two splice-site variants (affecting POLR1B and TADA2A genes), one frameshift (URB1), and one missense (PNKP) variant, resulting in a complete loss-of-function of these essential genes. The recessive lethal alleles affect up to 2.9% of the litters within a single population and are responsible for the death of 0.52% of the total population of embryos. Moreover, we provide compelling evidence that the identified embryonic lethal alleles contribute to the observed heterosis effect for fertility (i.e. larger litters in crossbred offspring). Together, this work marks specific recessive lethal variation describing its functional consequences at the molecular, phenotypic, and population level, providing a unique model to better understand fertility and heterosis in livestock.

Author summary

Lethal recessives are mutations that cause early lethality in homozygous state that usually occur at very low frequency in wild and domestic populations. In livestock, however, those mutations might become more prevalent as a result of inbreeding. In this study, we report five such recessive lethal haplotypes that cause embryonic lethality in homozygous state in pigs. The causal mutations are of different type but all destroy the structure of essential genes involved in cellular housekeeping processes, essential for embryonic development. The lethal recessives have substantial impact on the population fitness affecting up to 3% of the population litters, causing the death of 0.52% of the total population of embryos. Moreover, these 'natural knockouts' can increase understanding of gene function within the mammalian clade. Together, our study will allow monitoring, and facilitate the purging and partial elimination of recessive lethal mutations in frequently used pig breeds.

6.1 Introduction

Lethal recessive alleles cause pre- or postnatal death in homozygous affected individuals, reducing fertility in various populations (Cole et al. 2018). Although recessive lethals are generally widespread throughout populations, their effect is generally masked by the extremely low frequency of individual mutations. However, within small sized domestic and wild populations, those alleles might be exposed by inbreeding (Trask et al. 2016; Bosse et al. 2018), caused by matings between related parents that inherited the same recessive lethal allele from a common ancestor.

The precise impact of recessive lethals depends on the population structure (i.e. effective population size) and recessive lethal mutation rates. In livestock, populations have been subject to intensive (genomic) selection resulting in relative small effective population sizes (Hall 2016). With small effective population size, genetic drift can rapidly increase the frequency of recessive lethals in the population. Although genomic selection has enabled substantial improvement on various traits including production, fertility, and disease resistance (Gonzalez-Pena et al. 2015), it does not provide much advantage over traditional selection when it comes to controlling the frequency of recessive lethal mutations (Dalton et al. 2015).

Several studies have reported recessive lethal variation (i.e. death of embryo or foetus prior to birth), likely derived from a single sire origin, to be maintained in livestock populations (Derks et al. 2017; Cole et al. 2018). In fact, the frequency of some recessive lethals were driven by heterozygote advantage for important production traits, e.g. milk yield in cattle (Kadri et al. 2014), or growth in pigs (Derks et al. 2018), although the majority was likely the result of genetic drift. Together these studies show that lethal recessive alleles can have a considerable impact on population fitness, emphasizing the need for early detection. Although various recessive embryonic lethal loci have been reported in livestock, pinpointing the causal mutation can be extremely difficult. Charlier et al (2016) showed, using a reverse genetic screen, that loss-of-function mutations and deleterious missense mutations cause embryonic lethality in cattle populations. Nevertheless, the discovery of recessive embryonic lethals is often hampered by the lack of affected individuals and the relative low frequency. Genotyping and sequencing large cohorts of animals within single populations can therefore facilitate the discovery of such detrimental variation, and point directly to the causal mutations.

Pig fertility has increased steadily over the past years (Zak et al. 2017). Breeding for improved fertility concerns a large number of traits with a combined effect on overall fertility, and lethal recessives are increasingly considered to substantially affect fertility in purebred livestock populations (Casas and Kehrli 2016). However, in pigs, the final production animals are crossbreds between purebred populations, usually derived from three-way crosses (Hidalgo et al. 2016; Knol et al. 2016). First, crossbred sows are created from two elite purebred populations selected for high production of piglets (i.e. 'maternal lines'), which then are crossed with a third elite purebred population especially selected for meat production traits (i.e. 'paternal line'). These crossbreds are known to perform better on multiple traits compared to their parental purebred lines, in particular for traits related to fertility and robustness (Cassady et al. 2002), as a result of the heterosis effect. Heterosis is caused by different non-additive effects, such as dominance, and it has been subject to a scientific controversy; the dominance hypothesis emphasizes the suppression of undesirable recessive alleles (by dominant alleles), while the overdominance hypothesis emphasizes on heterozygote advantage (Charlesworth and Willis 2009). However, the magnitude of recessive lethals contributing to heterosis is largely unknown.

In this study we aim to explore the impact of lethal recessive variation in two pig populations using the following stepwise approach (1) perform simulations to assess the impact of genetic drift on lethal recessives, (2) identify haplotypes harboring lethal alleles using large-scale genotype data as developed by VanRaden et al. (VanRaden et al. 2011), (3) confirm lethality by reduced fertility in carrier animals, (4) identify causal mutations segregating on these haplotypes using whole genome sequence data (WGS) and RNA-sequencing data, (5) study the impact of recessive lethals on heterosis for fertility related traits.

6.2 Results

6.2.1 Population genetics of recessive lethal alleles

6.2.1.1 Estimating the number of recessive lethals segregating in two pig populations

In this study we analysed large-scale genomics, transcriptomics, and phenotype data from two commercial pig populations (Landrace and Duroc) to study recessive lethal alleles. We first evaluated the expected number and average frequency of recessive lethals within these two populations. Both the number and average frequency is a function of recessive lethal mutation rates and effective population size. The pig populations under study have an effective population size (Ne) in the range 100–150 (Hidalgo et al. 2016). Assuming similar mutation rates (~0.015 recessive lethals per gamete) as described for humans (Gao et al. 2015) and cattle (Charlier et al. 2016), we estimate that about 20 recessive lethal alleles are segregating (at average 2% allele frequency) in each of the pig populations under study. This corresponds to about one recessive lethal allele carried per individual and the death of 1% of the embryos in the population as a result of homozygosity for a recessive lethal allele (Charlier et al. 2016).

6.2.1.2 Simulating the impact of genetic drift on recessive lethals

The impact of genetic drift on recessive lethals heavily depends on the population structure and Ne. Small Ne leads to high extinction rates of *de novo* recessive lethals, but the few that are not lost tend to spread and increase in frequency. However, at a certain frequency, a trade-off between drift and selection is reached, at which the loss of homozygotes will prevent further increase assuming no heterozygote advantage. We evaluated this trade-off value (i.e. the maximum allele frequency reached by drift) using the actual population structure of the pig populations under study. We simulated the allele frequency change of a recessive lethal allele (fitness of homozygote mutants set to 0) over 25 generations in 1000 replicate populations with different start frequencies (S1 and S2 Figs**,** see Methods for details). Across simulations, the median frequency declines slightly with time (S1 Fig), but the decline is slower at lower allele frequencies (S2 Fig), due to very low number of carrier-by-carrier matings exposing the negative fitness effect. Interestingly, at about 10% allele frequency, the loss of homozygotes seems to prevent further increase of the allele frequency in the population (S1 and S2 Figs). This upper boundary is not observed under neutral assumptions (no negative fitness effect), in which the allele frequency can rise up to 30–40% within 25 generations (S3 Fig). Together, these results show that lethal alleles can reach allele frequencies up to 10% (20% carrier frequency) by genetic drift alone, although this happens only for a small fraction of the lethal variants.

In addition, we studied what proportion of the *de novo* recessive lethal mutations, is expected to remain in the population after 10 generations despite their very low starting frequency (0.024%). From the total number of *de novo* mutations, we show that about 2% still segregates after 10 generations (S4 Fig), and 1% after 25 generations (S5 Fig). We observe a similar pattern for neutral and recessive lethal *de novo* mutations, as there is very little purging efficiency at very low allele frequency (<2%).

6.2.2 Detection of haplotypes harbouring lethal recessives segregating at moderate frequencies in two purebred pig populations

To identify lethal alleles segregating in the pig populations we examined genotype data from 28,085 (Landrace), and 11,255 (Duroc) animals. All animals were genotyped or imputed to a medium-density 50K SNPchip (S1 Table). The genotypes were phased to build haplotypes, and then we applied an overlapping sliding window approach to identify haplotypes that show a deficit in homozygosity, likely harbouring a lethal recessive allele (VanRaden et al. 2011). The analysis yielded one strong candidate haplotype (DU1) harbouring a lethal recessive allele in the Duroc population, and four candidates in the Landrace population (LA1-4), respectively (Table 6.1). Haplotype lengths range from 0.5 to 5 Mb and carrier frequencies range from 4.6 to 13.4%. We observe no homozygotes for DU1, LA1, and LA3 haplotypes, while we expected 26, 126, and 16, respectively. We do observe two, and three homozygotes for LA2 (50 expected) and LA4 (14 expected), suggesting incomplete linkage disequilibrium (LD) of the haplotypes with the causal lethal recessive mutation. Four out of five haplotypes show deviation from Hardy-Weinberg equilibrium with over 50% carrier offspring for carrier-by-carrier matings. This is in concordance with the absence of homozygous offspring, resulting in a 1:2 offspring ratio instead of the expected 1:2:1 genotype offspring ratio (Figure 6.1A, Table 6.1).

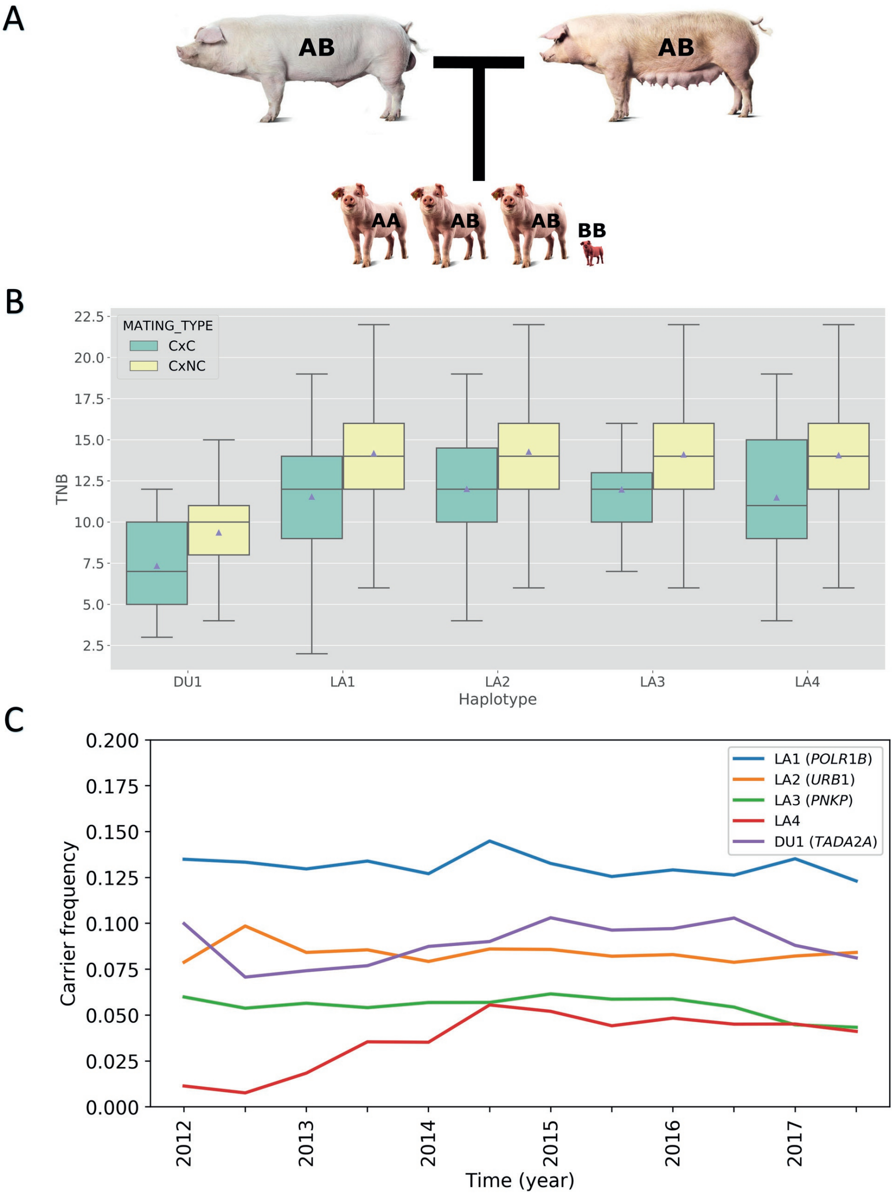


Figure 6.1 A) Example of a carrier-by-carrier mating in Landrace. CxC litters will result in a 1:2 genotype ratio instead of the normal 1:2:1 genotype ratio. **B)** Fertility phenotypes for lethal recessives. A significant reduction in total number born (TNB) is observed for CxC compared to CxNC matings. **C)** Carrier frequency for lethal alleles in the period 2012–2018. Figure shows relative stable carrier frequencies over a time period of 6 years (except for LA4).

Table 6.1 Haplotypes exhibiting missing or deficit homozygosity. Table shows five loci exhibiting missing or deficit homozygosity on the Sscrofa11.1 genome build, four in the Landrace population (LA1-4), and one in the Duroc population (DU1). The table shows the genomic location, carrier frequency, and deficit of homozygosity for each haplotype. The deficit of homozygosity is calculated based on trio information (parents and offspring) with the formula described by Fritz et al., 2013 [20], and from haplotype frequency, using the Hardy-Weinberg principle. Genotyped progeny is derived from CxC matings.

Hap.	SSC	Start	End	#Carriers	Carrier Freq.	Expected (trio)	Expected (freq.)	Obs.	Exact binomial test	# Genotyped progeny	# Het. progeny
DU1	12	38.5	39.0	1,084	9.6	7 .0	26.1	0	1.81e-13	28	18 (64.3%)
LA1	3	42.6	47.5	3,763	13.4	52.0	126.0	0	2.11e-63	208	120 (57.7%)
LA2	13	195.7	196.2	2,358	8.4	18.25	49.5	2	6.39e-22	73	53 (72.6%)
LA3	6	52.5	54.0	1,319	4.7	6.0	15.5	0	2.54e-08	24	11 (45.8%)
LA4	12	25.0	27.0	1,271	4.6	-	14.4	3	0.00017	5	3 (60%)

6.2.3 Carrier-by-carrier matings produce significantly smaller litters

We analysed the effect of the haplotypes on fertility phenotypes including total number born (TNB), number born alive (NBA), number of stillborn (NSB), and number of mummified piglets (MUM). We examined a total of 504 carrier-by-carrier (CxC) and 5,992 carrier-by-noncarrier (CxNC) matings (Table 6.2). Interestingly, all five haplotypes show significant reduction in both TNB (Table 6.2, Figure 6.1B) and NBA for CxC matings (S2 Table). The reduction in TNB ranges from 15.1 to 21.6% which is somewhat smaller than the expected 25% assuming early lethality with complete penetrance for homozygotes (Table 6.2). No significant increase in number of stillborn (NSB) or mummified piglets (MUM) was found, suggesting that homozygotes die very early in pregnancy (S3 Table). Together the 504 CxC matings cause a loss of 1,261 piglets over the last 5 years (comparing average litter size of CxC and CxNC matings), affecting 2.9% and 0.92% of all litters in the Landrace and Duroc population, respectively (Table 6.2). None of the five regions were previously reported to be associated with reduced TNB (Hu et al. 2016).

6.2.4 Candidate embryonic lethal alleles predicted from whole-genome sequence (WGS) and RNA-sequencing (RNA-seq) data

To find causal mutations, we analysed WGS (Landrace: 167, Duroc: 119) and RNA-seq (Landrace: 34, Duroc: 25) data available from the populations under study (S4–S5 Tables). The data was mapped to the latest Sscrofa11.1 reference build and functionally annotated using the Variant Effect Predictor (VEP) (McLaren et al. 2016).

Next, we focused on variants likely causing embryonic lethality (EL) in homozygous state, examining the impact of individual variants on the proteins. First, we selected loss-of-function (LoF) variants (frameshift, stop-gained, splice-site) and predicted deleterious missense variants within each population (Kumar et al. 2009). The predicted LoF and deleterious mutations show clear patterns of purifying selection, as observed from generally lower allele frequencies (S6 Fig), an enrichment of inframe indels (S7 Fig), and an enrichment of LoF mutations in the N-and C-terminal end of the gene (S8 Fig).

Table 6.2 Fertility phenotypes for total number born. Table shows the number of CxC and CxNC mating for each haplotype, the reduction in total number born (TNB), the percentage of affected litters in the population, the piglet loss associated with the CxC matings, the percentage of embryo deaths in the entire population, and the overall population piglet reduction.

Population	Hap.	#CxC	#CxNC	TNB (CxC)	TNB (CxNC)	Reduction	% Affected litters	Piglet loss	% Death	Population piglet reduction*
Landrace	LA1	297	2,350	11.51	14.18	18.8%	1.796	792.99	0.338	0.0479
	LA2	127	1,527	12.00	14.26	15.9%	0.706	287.02	0.112	0.0159
	LA3	30	872	11.96	14.09	15.1%	0.212	63.90	0.032	0.0045
	LA4	29	950	11.48	14.05	18.3%	0.212	74.53	0.039	0.0055
	SUM	**483**	**5,699**	-	-	-	**2.926**	**1218.44**	**0.521**	**0.0739**
Duroc	DU1	21	293	7.33	9.35	21.6%	0.922	42.42	0.199	0.0186

* Calculated as the product of the average TNB (Landrace: 14.18, Duroc: 9.35) and the population deaths in the Landrace and Duroc population.

6.2.5 Identifying candidate LoF mutations in lethal haplotypes

The likelihood of carriers being present in even a small random sample of pigs is high due to the relatively high carrier frequency of the candidate haplotypes found in this study. The candidate haplotypes could therefore be identified in pigs of the same populations for which WGS data (LA1: 21, LA2: 17, LA3: 7, LA4: 9, DU1: 9) or RNA-seq data (LA1: 4, LA2: 3, DU1: 3) was available. For each of the five haplotypes we used criteria of physical distance and co-segregation (see Methods for details) to select candidate causal mutations. A single strong candidate mutation was identified for all haplotypes, except LA4 (Table 6.3).

Table 6.3 Candidate causal variants for lethal haplotypes. The table shows the type, location, the affected gene, and the predicted impact for each candidate recessive lethal variant. The relative position in the protein shows the position of the variant relative to the protein length, for splice-variants, the affected intron is presented.

Hap.	Type	SSC	Position	Ref	Alt	Gene	AA change	Relative pos. in protein	Gene name
DU1	Splice-donor	12	38,922,102	G	A	*TADA2A*	p.Ile319fs	Intron 13	Transcriptional adaptor 2A
LA1	Splice-region	3	43,952,776	T	G	*POLR1B*	p.Ile701fs	Intron 14	RNA polymerase I subunit B
LA2	Frameshift	13	195,977,038	C	-	*URB1*	p.Val1961fs	0.87	Ribosome biogenesis homolog
LA3	Missense	6	54,880,241	T	C	*PNKP*	p.Gln96Arg	0.17	Polynucleotide kinase 3'-phosphatase

6.2.5.1 A splice donor mutation in TADA2A induces embryonic lethality in Duroc (DU1 haplotype)

Whole genome sequence data from nine DU1 carrier animals revealed 20 variants in high LD ($r^2 > 0.8$) with the DU1 haplotype (S6 Table), of which only one variant is predicted to have high impact. This variant, a heterozygous splice-donor mutation (12:g.38922102G>A) in the Transcriptional adapter-Ada2 (*TADA2A*) gene is in complete LD with the DU1 haplotype (Table 6.3). The mutation affects a conserved GT splice dinucleotide site at the 5’ end of the intron between exons 13 and 14 (S9 and S10 Figs). We evaluated the effect on RNA splicing using RNA-seq data from three carrier animals (S5 Table). The splice-donor mutation seems to cause retention of intron 13 between exon 13–14 in one of the samples (S11 Fig), shown by reads spanning the exon-intron boundaries on the splice donor and acceptor sites in intron 13, not seen for non-carriers. Interestingly, two other carrier samples show exon skipping of exon 13 (S9 Fig), resulting in a frameshift, the addition of a novel methionine, and a premature stop codon in the first codon of exon 14. The mutant mRNA codes for a truncated TADA2A protein (318 amino acids) lacking the terminal 101 amino acids (AA) that includes the conserved SWIRM domain required for DNA binding (Qian et al. 2005). These results show that the splice-donor mutation affects *TADA2A* splicing with different consequences (both exon skipping and intron retention) in carrier animals, but in all cases result in a compromised, non-functional transcript. The TADA2A protein is involved in the general transcription machinery and it’s gene is known to be essential in yeast and drosophila (Pankotai et al. 2005).

However, no information for mice null-mutants is available for *TADA2A* (Blake et al. 2017).

6.2.5.2 A splice region mutation in POLR1B induces early embryonic lethality in Landrace (LA1 haplotype)

The LA1 haplotype was the longest haplotype observed in this study (SSC3:42.6–47.5). Therefore, we first performed a fine-mapping analysis to further pinpoint the region containing the causal mutation. We observed two recombinant animals (S7–S8 Tables) that were homozygous for a part of the LA1 haplotype in the region (45.6–47.5), leaving a final candidate region of length 3Mb (SSC3: 42.6–45.6 Mb). Whole genome sequence data from twenty-one LA1 carrier animals revealed a set of 415 variants, and one small intronic deletion in high LD ($r^2 > 0.8$) with the LA1 haplotype (S9 Table), of which five variants are located within coding sequence (2 missense, 1 synonymous) or splice regions (2 splice-region). Both missense variants are predicted to be tolerated by SIFT, unlikely to be causal (S9 Table). However, the splice region mutation in intron 14 of the RNA polymerase I subunit B (POLR1B) gene is predicted to have high impact (Table 6.3, S12 Fig). The splice mutation affects a conserved adenine in the GTRAG splice site motif (positive strand: 3:g.43952776T>G, Figure 6.2A and 2B). The adenine is conserved throughout a wide range of vertebrate species (S12 Fig). Next, we analysed the RNA-seq data from four carrier animals and found that the splice region mutation causes exon skipping of exon 14 in all four carrier animals (Fig 6.2C, S13 and S14 Figs), not observed for non-carrier animals (S15 Fig). POLR1B isoforms that show alternative splicing for exon 14 have not been annotated in pigs or any other mammals, including human and bovine embryonic tissues. Skipping of exon 14 introduces a glutamic acid and a premature stop codon in the second codon of the terminal exon, lacking the final 370 amino acids located in the conserved subunit 2, hybrid-binding domain (binding to the DNA strand) (Fig 6.2D). Hence, this splice-region mutation likely causes a complete LoF of the POLR1B protein. The structure of RNA-polymerase 1, and the affected POLR1B subunit is presented in S16 Fig. The POLR1B gene is strongly conserved among vertebrates and null-mutant mice show embryonic lethality even prior to implantation (Chen et al. 2008).

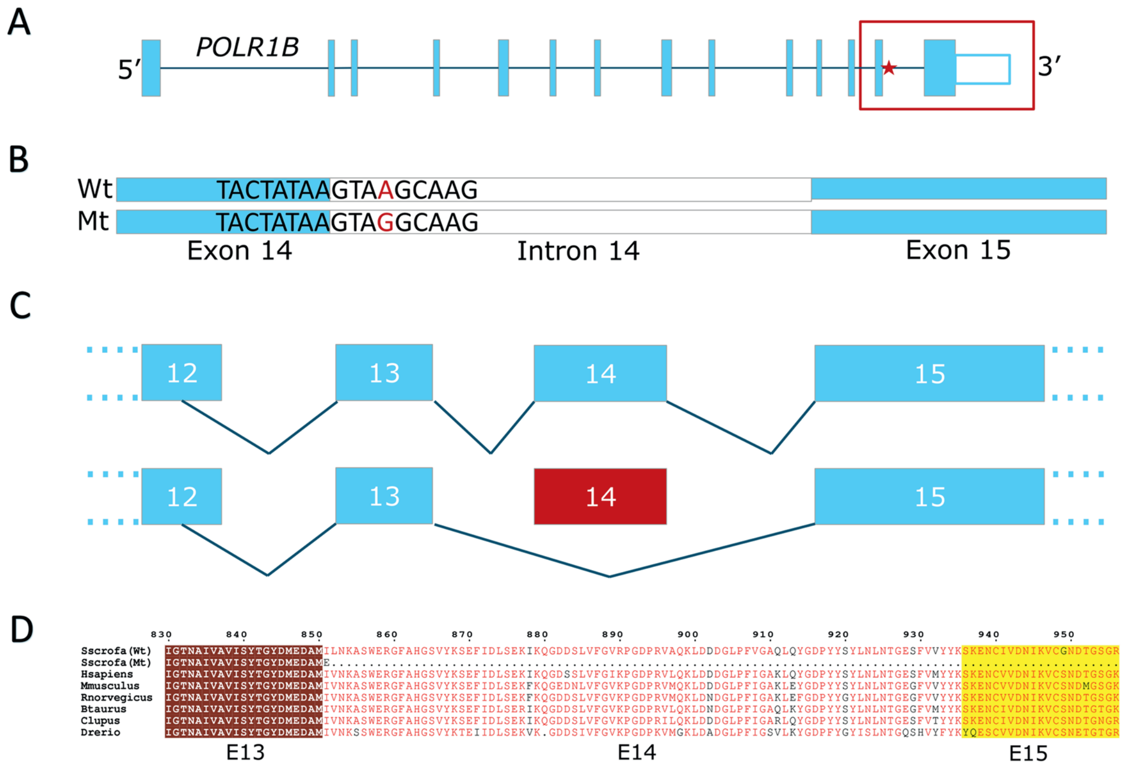

A
POLR1B
5′
3′
B
Wt
TACTATAAGTAAGCAAG
Mt
TACTATAAGTAGGCAAG
Exon 14
Intron 14
Exon 15
C
12
13
14
15
12
13
14
15
D
830 840 850 860 870 880 890 900 910 920 930 940 950
Sscrofa(Wt) IGTNAIVAVISYTGYDMEDAMILNKASWERGFAHGSVYKSEFIDLSEKIKQGDDSLVFGVRPGDPRVAQKLDDDGLPFVGAQLQYGDPYYSYLNLNTGESFVVYYKSKENCIVDNIKVCGNDTGSGR
Sscrofa(Mt) IGTNAIVAVISYTGYDMEDAME...
Hsapiens IGTNAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSEKIKQGDSSLVFGIKPGDPRVLQKLDDDGLPFIGAKLQYGDPYYSYLNLNTGESFVMYYKSKENCVVDNIKVCSNDTGSGK
Mmusculus IGTNAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSEKFKQGEDNLVFGVKPGDPRVMQKLDDDGLPFIGAKLEYGDPYYSYLNLNTGEGFVVYYKSKENCVVDNIKVCSNDMGSGK
Rnorvegicus IGTNAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSEKFKQGDDSLVFGVKPGDPRVMQKLDNDGLPFIGAKLEFGDPYYGYLNLNTGEGFVVYYKSKENCVVDNIKVCSNDTGSGK
Btaurus IGTNAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSERIKQGDDSLVFGVKPGDPRVLQKLDNDGLPFIGAQLKYGDPYYSYLNLNTGEGFVMYYKSKENCIVDNIKVCSNDTGTGK
Clupus IGTNAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSEKIKQGDDSLVFGVKPGDPRILQKLDDDGLPFIGARLQYGDPYYSYLNLNTGESFVTYYKSKENCIVDNIKVCSNDTGNGR
Drerio IGTNAIVAVISYTGYDMEDAMIVNKSSWERGFAHGSVYKTEIIDLSEKVK.GDDSIVFGVKPGDPKVMGKLDADGLPFIGSVLKYGDPFYGYISLNTGQSHVYFYKYQESCVVDNIKVCSNETGTGR
E13
E14
E15

Figure 6.2 A) *POLR1B* gene model. The location of the mutation on the splice-donor site of intron 14 is indicated with a red star. **B)** Illustration of the affected exon-intron splice region. The causal 3:g.43952776T>G mutation is indicated in red. **C)** Exon skipping of *POLR1B*. The mutation causes complete exon skipping of exon 14, resulting in a truncated mRNA. **D)** Alignment of the mutant (Mt) and wildtype (Wt) POLR1B protein sequence. Skipping of exon 14 introduces a glutamic acid and a premature stop codon in the second codon of the terminal exon.

6.2.5.3 A frameshift mutation in URB1 causes embryonic lethality in Landrace (LA2 haplotype)

Whole genome sequence data from seventeen LA2 carrier animals revealed a set of 234 variants and one small intronic deletion in high LD ($r^2 > 0.8$) with the LA2 haplotype (S10 Table), of which five variants are located within coding sequence (1 frameshift, 1 missense, 2 synonymous) or splice regions (1 splice-region). The missense and splice-region variant in the ISTN1 gene are predicted to be tolerated, unlikely to be causal (S10 Table). However, the frameshift mutation in exon 38 of the *URB1* gene (13:g.195977038delC) caused by a 1-bp deletion is predicted to have high impact (Table 6.3, S17 Fig). The frameshift (ENSSSCP00000036505:p.Val1961fs) introduces 26 novel amino acids and a premature stop codon, producing a truncated protein of 1,986 amino acids, lacking the final 261 amino acids compared to the wild-type protein (2,247 AA). *URB1* (Ribosome Biogenesis 1 Homolog) is involved in the biogenesis of the 60S ribosomal subunit and is an essential gene in yeast and drosophila (Giaever et al. 2002), but no information from mouse null-mutants is available for this gene. Moreover, no homozygous LoF mutations are reported in the ExAc database for the *URB1* gene (Lek et al. 2016), supporting that a functional copy is required.

6.2.5.4 A missense mutation in PNKP is a candidate to cause embryonic lethality in Landrace (LA3 haplotype)

Only four variants are found to be in high LD with the LA3 haplotype (S11 Table) including one deleterious missense mutation in the *PNKP* gene (6:g.54880241G>T), predicted to be strongly deleterious by SIFT (0.02) and PROVEAN (-2.9). The missense mutation causes a glutamine to arginine amino acid substitution (ENSSSCP00000003467:p.Gln96Arg) (Table 6.3, S18 Fig). The glutamine residue is highly conserved among vertebrates (S19 Fig), and is part of the protein-protein interaction FHA and SMAD domain. The *PNKP* gene plays a key role in the repair of DNA damage, being an essential part in the non-homologous end-joining (NHEJ) and base excision repair (BER) pathways. Mouse null-mutants exhibit embryonic lethality (Shimada et al. 2015). However, homozygous loss-of-function mutations in PNKP are associated with various neurologic diseases in human, but not with early lethality (Dumitrache and McKinnon 2017).

6.2.6 Nonsense mediated decay of alternatively spliced transcripts

We assessed whether the splice mutations in *TADA2A* and *POLR1B* are subject to nonsense-mediated mRNA decay, a surveillance pathway eliminating transcripts that contain premature stop-codons (Hug et al. 2016). We assessed the expression of the wild-type and mutant transcripts in carrier animals for both genes. The abundance of both the mutant *TADA2A*, and *POLR1B* transcripts are significantly lower (2.5- to 5 fold) compared to the wild-type transcripts, supporting that the mutant transcripts are likely subject to nonsense mediated decay (S12 Table).

6.2.7 Validation of candidate causal mutations in carrier-by-carrier litters

We genotyped the complete litters of three LA1, and one LA2 CxC mating for the predicted causal mutations, and confirmed the carrier status of both parents for each litter (S13–S14 Tables). The three LA1 litters produced 38 piglets, 14 were homozygous for the wild-type allele (36.8%), 24 were heterozygous carriers (63.2%), and no homozygous mutants were found (P<0.005, Table 6.4). The LA2 litter produced 13 piglets, 3 homozygous wild-type, 10 heterozygous carriers of the deletion, and no homozygous del/del mutants (P = 0.076). These results are in line with the 1:2 genotype ratio expected for CxC litters, supporting the recessive lethality of the candidate causal mutations.

Table 6.4 Genotyping of causal mutations in four carrier by carrier litters. The parents (sow and boar) and complete liveborn and stillborn progeny are genotyped for the candidate causal mutations. Table shows the number of progeny, type of birth, and genotypes for the four examined litters.

LitterID	Haplotype-Gene	Gene-Mutation	# Progeny	# Liveborn	# Stillborn	# Wt	# Carrier	# Lethal	*p* (Chi-Square)
1	LA1-*POLR1B*	*POLR1B* 3:g.43952776T>G	14	13	1	TT = 4	TG = 10	GG = 0	
2	LA1-*POLR1B*	*POLR1B* 3:g.43952776T>G	11	11	0	TT = 3	TG = 8	GG = 0	
3	LA1-*POLR1B*	*POLR1B* 3:g.43952776T>G	13	12	1	TT = 7	TG = 6	GG = 0	
		SUM—LA1	**38**	**36**	**2**	**TT = 14**	**TG = 24**	**GG = 0**	**p<0.005**
4	LA2-*URB1*	*URB1* 13:g.195977038delC	13	11	2	CC = 3	C/Del = 10	Del/Del = 0	p = 0.076

6.2.8 Embryonic lethal alleles are generally maintained at stable population frequencies

The current frequency of the embryonic lethals raises the question how population frequencies of lethal alleles have developed over the past years. Interestingly, we observe that overall the recessive lethal alleles are maintained at relative stable frequency over the past seven years (2012–2018, Figure 6.1C), despite ongoing selection on littersize in these populations.

6.2.9 Large-scale phenotype data supports both balancing selection and genetic drift driving the frequency of recessive lethals

High frequency of a lethal allele can be caused by a trade-off between a negative trait (i.e. reduced fertility) and another trait, e.g. improved growth (Derks et al. 2018). We tested whether carriers of lethal haplotypes show signs of heterozygote advantage on one of the traits included in the breeding goal, which could potentially drive the frequency of the allele. However, we only found strong association signals for the most frequent haplotype LA1 (S15 Table). The carriers of the haplotype LA1, compared to non-carriers, show: increased mothering ability (fewer piglet deaths, and larger piglet weight at 21 days), increased carcass quality (larger loin depth, less backfat, and higher meat percentage), lower meat quality (less intramuscular fat), and slower growth (S15 Table). As expected, only a small negative effect on total number born is observed, as reduced litters will only be expressed in CxC matings. Mothering ability (consisting of several maternal traits) is a very important part of the breeding goal for this breed, and could be a strong candidate to support heterozygote advantage for this group of traits. Sows that carry the LA1 haplotype show lower piglet mortality at 21 days, and increased piglet weight at 21 days. For the other haplotypes, we only obtain weak associations (S16 Table), suggesting that the current frequency of the haplotypes are likely the result of genetic drift rather than heterozygote advantage.

6.2.10 Lethal recessives explain part of the heterosis effect on fertility traits in the crossbred litters

The impact of individual lethal recessive alleles largely depends on its frequency. For example, assuming random matings, we estimate that about 1.8% of the population litters are CxC matings for LA1, while only 0.21% of the litters are CxC matings for the LA4 allele (Table 6.2). The four Landrace lethals combined affect 2.9% of the litters within the population, responsible for the death of 0.52% of the total population of embryos, which causes an average reduction of 0.073 TNB in the population (Table 6.2). Next, we investigated whether ELs could contribute to the heterosis effect for fertility (TNB) observed in the crossbreds. The current Landrace population is mostly crossed with a Large White (LW) population to generate a commercial F1 population. F1 litters in Landrace sows produce on average 0.20 piglet larger litters compared to purebred Landrace (LR = 14.18, LR/LW = 14.38) (S17 Table). All three identified mutations (LA1-LA3) are not segregating in the LW population, suggesting no homozygous affected individuals in the F1 population. Therefore, part of the TNB difference is likely caused by the four recessive lethals affecting the purebred litters (given the average reduction of 0.073 TNB as a result of the four ELs). Nevertheless, other heterotic effects will contribute to the increased litter size as well.

6.3 Discussion

In this study we report five embryonic lethal haplotypes that segregate with carrier frequencies in the range of 4.6–13.4% in two commercial pig populations. We show that the use of large-scale genotype data within single populations provides the power to find lethal alleles with low frequencies. The inclusion of over 28 thousand individuals from the Landrace population, for instance, allowed us to detect the LA4 haplotype that has an allele frequency of only 2.3%. For three of the five recessive lethal haplotypes no homozygous carrier individuals were found, suggesting complete LD with a causal, recessive lethal variant. However, none of the recessive embryonic lethal haplotypes resulted in the theoretically expected 25% reduction in piglets born (range observed is: 15.1–21.6%). The most likely explanation is that the number of embryos frequently exceeds the uterine capacity of the sow. Hence, by reducing the number of embryos by 25%, fewer wildtype/wildtype and wildtype/mutant embryos are eliminated (Da Silva et al. 2017). This compensatory effect could be especially relevant if homozygous affected zygotes fail to develop or embryos die very early on. Especially if they die prior to implantation in the uterus, other viable healthy zygotes can compensate (i.e. take their place in the uterus) of the lethal effect in homozygous zygotes. A compensatory effect is particularly likely for LA1 homozygous affected embryos (*POLR1B*), since in homozygous *POLR1B* knock-out mice embryos terminate development before implantation in the uterus is established (Chen et al. 2008). Moreover, we did not observe an increase in mummified or stillborn piglets for CxC matings, again suggesting early termination (i.e. prior to day 35 in gestation) of homozygous animals in utero.

All four genes affected by embryonic lethal alleles are involved in cellular housekeeping functions including transcription (*POLR1B*, *TADA2A*), translation (*URB1*), and DNA damage repair (*PNKP*), supported by the relative high expression of these genes within different tissue types (Freeman et al. 2012). The RNA-seq data from carrier animals confirmed the functional impact of the DU1 splice-donor and LA1 splice-site mutations, both resulting in truncated proteins caused by the skipping of complete exons. Interestingly, we show that a single splice-donor mutation can simultaneously cause exon skipping and intron retention, something described previously in human studies (Kallabi et al. 2015), but not previously observed in pigs. Moreover, the alternatively spliced mRNAs are likely subject to nonsense-mediated decay, because the level of the mutant mRNA is significantly lower compared to the wild-type mRNA. All mutations (except *URB1*) are located within parts of the genes predicted to be intolerant to LoF mutations observed from a negative subRVIS score (Gussow et al. 2016). Interestingly, embryonic lethality has been described in targeted mice null-mutants for *POLR1B* and *PNKP* (Chen et al. 2008; Shimada et al. 2015), but not for *URB1* and *TADA2A*. In this study, however, we demonstrate that both *URB1* and *TADA2A* are essential for normal embryonic development in pigs, likely to be similar in human. We did not find any coding variants or structural variants that are in high LD with the LA4 haplotype (S18 Table). However, other type of variants (e.g. small insertion elements) could also induce embryonic lethality or genetic disease (Schutz et al. 2016), something not well explored in this study.

We show that the frequency of the lethal haplotypes over time, at least over the past seven years, is stable, suggesting that there is no strong selection against these recessive lethal variants. The population genetic analysis indicates that the observed frequencies of recessive lethal alleles found in this study, can be the result of genetic drift alone. Moreover, the study on *de novo* mutations shows that the lethal mutations in the LA1, LA2, and DU1 haplotypes (allele frequency > 4%) likely arose over 25 generations ago (assuming no heterozygote advantage). Genetic drift as a driving force for the observed frequencies is further supported by the lack of clear evidence for heterozygote advantage (except for LA1). Nevertheless, we cannot exclude that these alleles have been subject to genetic-hitchhiking in the past, resulting in heterozygote advantage due to LD with a beneficial allele that became fixed.

Evidence for heterozygote advantage has been found for other highly detrimental variants that occur in higher frequencies than those observed here in wild and domesticated populations (Kadri et al. 2014; Derks et al. 2018). This could also be the case for the most frequent recessive lethal in our study, LA1, for which a highly positive effect on mothering ability for heterozygous carriers was found. In sow lines, mothering abilities are among the most important selection traits. The favorable phenotype of heterozygous carriers (mothering ability) offsets the occasional lower litter size, as long as the carrier frequency does not become too high. Nevertheless, our simulations show that the allele frequency of recessive lethals can rise up to 10% as a result of genetic drift alone. At this frequency, the negative effects on fitness

(i.e. smaller litters and lack of homozygotes) will prevent further increase in allele frequency.
Recessive lethals, by definition, deviate from the Hardy-Weinberg equilibrium (HWE). We analyzed whether our liberal HWE marker threshold might hampered the detection of high frequency ELs, but no new high frequency haplotypes were revealed (S21 Table). Nevertheless, not all embryonic lethal variation currently present in the populations under study was identified. In fact, even if LD between recessive lethal causal variants and SNP-chip based haplotypes would be perfect, the minimum allele frequency that could be detected is around 2% for the Landrace population, and around 4% for the Duroc population. In addition, lethal recessives residing on more common haplotypes cannot be detected because the SNP density is likely too low to distinguish between the haplotype with and the haplotype without the lethal recessive. We estimated that ELs likely account for 1% of deaths in these pig populations, but the four identified Landrace lethals account for the loss of 0.52% of all newborn pigs per generation, showing that the remainder 0.48% is caused by yet to be identified EL mutations.

In pigs, the crossbred production animals show clear signs of heterosis, especially for fertility related traits (Cassady et al. 2002). We provide compelling evidence that embryonic lethals contribute to the heterosis effect seen in the Landrace crossbred litters. Assuming that recessive lethal variation is generally occurring in a single breeding line only, crossbred products will only be heterozygous for the lethal recessive mutations. We show that at least 2.9% of the litters within a single pure breeding line (Landrace) are offspring of matings between carriers of lethal recessives identified in this study, and that the four identified lethal variants are responsible for a significant part of the total heterosis effect (as measured in surviving piglets). The heterosis effect is caused by the suppression of recessive lethal alleles by dominant wildtype alleles in the crossbreds (Charlesworth and Willis 2009), providing evidence that the impact of lethal recessives on fertility and heterosis in these commercial pig populations is likely underestimated. Nevertheless, other detrimental, but not lethal alleles, uniquely segregating in purebred pig populations likely contribute to the heterosis effect even more, although this has never been properly quantified.

Our study shows high resolution and efficiency of combining large-scale genotype (SNP chip), phenotype, whole-genome sequence, and RNA-sequencing data to identify deleterious mutations that confer early embryonic lethality in pigs. We report five relatively common embryonic lethal alleles with carrier frequencies between 4.7–13.4%. Four of the variants destroy the structure of essential genes involved in cellular housekeeping processes including mRNA transcription, translation, and DNA repair. Simulation shows that observed allele frequencies can be mainly explained as consequence of drift only and there is no clear evidence for heterozygote advantage for favourable traits. The large amount of phenotype and genotype data collected in modern breeding programs in combination with increasing genomic data provides excellent possibilities to monitor old and new detrimental mutations segregating in purebred livestock populations. Although, we provide compelling evidence that the identified embryonic lethal alleles contribute to the observed heterosis effect for fertility, only a small proportion of the overall heterosis can be explained by the effect of the EL alleles detected. Other factors contributing to heterosis remain to be detected.

6.4 Methods

6.4.1 Ethics statement

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

6.4.2 Animals, genotypes and pre-processing

The dataset consists of 28,085 and 11,255 animals from Norwegian Landrace and Duroc purebreds, respectively. The animals are genotyped on the (Illumina) Geneseek custom 50K SNP chip with 50,689 SNPs (50K) (Lincoln, NE, USA). The chromosomal positions were determined based on the Sscrofa11.1 reference assembly. SNPs located on autosomal chromosomes were kept for further analysis. Next, the SNPs were filtered using following requirements: Each marker had a MAF greater than 0.01, and a call rate greater than 0.85, and an animal call rate > 0.7. SNPs with a p-value below $1x10^{-5}$ for the Hardy-Weinberg equilibrium exact test were also discarded. All pre-processing steps were performed using Plink v1.90b3.30 (Purcell et al. 2007). After quality control, the final dataset contained 43,375 and 42,706 markers for Landrace and Duroc populations, respectively.

6.4.3 Phasing and identification of missing homozygote haplotypes

We used BEAGLE version 4.1 genetic analysis software to phase both populations separately (Browning and Browning 2007). Haplotypes exhibiting missing or deficit homozygosity were identified using an overlapping sliding window approach from 0.5 to 5 MB. Within each window individual haplotypes (with a frequency > 0.5%) were evaluated for missing or deficit homozygosity. The expected number of homozygotes was estimated using two methods: (1) Estimation based on haplotype frequency, using the Hardy-Weinberg principle, (2) Estimation based on haplotype information from both parental haplotypes with the formula described by Fritz et al., 2013 (Fritz et al. 2013). An exact binomial test was applied to test the number of observed homozygotes with the number of expected homozygotes. Haplotypes were considered significant if $P < 5\times10^{-3}$.

6.4.4 Fine mapping of the LA1 haplotype

We examined the wild-type haplotypes for each LA1 carrier animal to identify recombinant individuals. We used PyVCF (Casbon 2012) to gather both haplotypes for all carriers animals within the LA1 genomic region from the BEAGLE phased VCF file. Next, we divided the LA1 haplotype in 5 shorter sub-haplotypes (length = 1Mb). Next, we examined whether the sub-haplotypes were carried in homozygous state in the group of LA1 carrier animals. Homozygous sub-haplotypes were excluded to carry the causal mutation.

6.4.5 Phenotypic effects associated with lethal haplotypes

We examined phenotypic records for TNB, NBA, NSB, and MUM to verify the lethality of the detected haplotypes. We listed all CxC and CxNC matings available and used a Welch t-test to assess if the phenotypes from the CxC matings significantly differ from CxNC matings. A *P*-value < 0.05 was considered significant. The order of CxNC matings does not reflect the sex of the parent animal and is both carrier boar and carrier sow combined

6.4.6 Population sequencing and mapping

Sequence data was available for 167 (Landrace) and 119 (Duroc) animals from paired-end 100 bp reads sequenced on Illumina HiSeq (van Son et al. 2017a). The sequenced samples are frequently used boars born between 2003 and 2017, selected to capture as much of the genetic variation present in the Landrace and Duroc populations. The majority of the sequenced animals were also represented in the 50K genotype dataset (Landrace = 161, Duroc = 72). The coverage ranges from 6.65 to 21.46, with an average coverage of 12.70 (S19 Table). Sickle software was used to trim the sequences. BWA-MEM (version 0.7.15, (Li and Durbin 2009)) was used to map the WGS data to the Sscrofa11.1 reference genome. Samtools dedup was used to discard PCR duplicates (Li et al. 2009). GATK IndelRealigner was used to perform local realignments of reads around indels (McKenna et al. 2010b).

6.4.7 Variant discovery

Freebayes variant calling software was used to call variants with following settings: min-base-quality 10—min-alternate-fraction 0.2—haplotype-length 0—ploidy 2—min-alternate-count 2 (Garrison and Marth 2012). Post processing was performed using bcftools (Li et al. 2009). Variants with low phred quality score (<20), low call rate (<0.7) and variants within 3 bp of an indel are discarded. Next, genotype calls are filtered for sample depth (min: 4, max: AvgDepth *2.5) leaving a total of 18,118,052, and 15,857,077 post-filtering variants for Landrace and Duroc population, respectively. The average variant call rate is 95.4% (Landrace) and 96.4% (Duroc), and the average transition / transversion (TS/ TV) ratio is 2.42 and 2.27, respectively, in concordance with previous findings in pigs (Bianco et al. 2015).

6.4.8 Structural variation

The Smoove pipeline (https://github.com/brentp/smoove) was used to call SVs. Smoove uses various software to call and filter SVs taking the alignment BAM files, and the Sscrofa11.1 reference genome as input. First, Lumpy software is used to call SVs (Layer et al. 2014). Next, Svtyper is used to genotype SVs (Chiang et al. 2015). To further filter SV calls, Mosdepth is used to remove high coverage regions, and Duphold to annotate depth changes within and on the breakpoints of SVs.

6.4.9 Functional annotation of variants

We performed variant (SNPs, Indels, and SVs) annotation using Variant Effect Predictor (VEP, release 90) (McLaren et al. 2016). The variant effect prediction in protein altering variants was performed using SIFT (Kumar et al. 2009). The following variant classes were considered potentially causing LoF: splice acceptor, splice donor, inframe indels, frameshift, stop loss, stop gained, and start lost variants.

6.4.10 Candidate embryonic lethal alleles

LoF and deleterious missense variants were selected within each population that met following criteria. The variant is found in a maximum of 1 homozygous individual, allowing one false genotype assignment. Next, the variant is annotated in a gene that is a 1-to-1 ortholog with cattle to minimize the effect of off-site mapping of sequence reads, which can be particularly problematic for large gene families. Finally, the list of EL candidates was manually validated for possible sequencing and alignment artefacts. Further functional support was obtained from the MGI database release 6.10 (i.e. phenotypes from null-mutant mice) to predict the relative impact on the phenotype (Blake et al. 2017). To identify candidate causal mutations for the haplotypes exhibiting missing homozygosity we applied the following criteria: 1) The mutation is located within 5 Mb of the haplotype boundaries. 2) The mutation is carried in heterozygote state by the haplotype carriers and no homozygous individuals are observed. 3) The mutation is absent from non-haplotype-carrier animals. 4). The mutation is in high LD with the candidate lethal recessive haplotype ($R^2 > 0.7$). LD analysis was performed using Plink v1.90b3.30 (Purcell et al. 2007) with following settings:—chr-set 18,—r2, ld-window-r2 0.7.

6.4.11 RNA sequencing and nonsense mediated decay

The impact of the splice mutations on the expression of the gene was assessed using RNA-seq data. The animals sequenced are frequently used artificial insemination boars selected based on extreme phenotypes all present in the genotyping data (van Son et al. 2017b). The phenotypes are based on high and low sperm DNA fragmentation index, a measure of well packed double-stranded DNA vs single-stranded denatured DNA, which is an important indicator of boar fertility. We mapped the RNA-seq data to the Sscrofa11.1 reference genome using STAR (Dobin et al. 2013) and called transcripts and FPKM expression levels using Cufflinks (Trapnell et al. 2012). To test for nonsense mediated decay, we examined the transcript expression level of both the mutant and wild-type transcript identified by Cufflinks. The predicted effect on the mRNA was further evaluated by manually inspecting alignments using the JBrowse visualization software (Skinner et al. 2009). Variants were called on the RNA using Freebayes v1.1.0 (Garrison and Marth 2012) to examine if the genes are subject to genomic imprinting, heterozygous coding variants are listed in S20 Table.

6.4.12 Validation of candidate causal mutations in carrier-by-carrier litters

We tracked four recent CxC litters and sampled the complete litter including parent animals. The complete litter and parents were genotyped for the candidate causal variants using matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (MALDI-TOF MS) assays. The candidate mutations were fitted into the same assay and the assay was designed using MassARRAY Assay Design software (Agena Biosciences, Hamburg, Germany). The genotyping was done using the IPLEX protocol according to manufacturer’s instructions. The difference in the expected and observed Mendelian genotype ratios was tested using a Chi-Square test.

6.4.13 Frequency and impact of embryonic lethal alleles

We analyzed the frequency of the haplotypes harboring embryonic lethals per half-year starting from 01-jan-2012 and assessed the frequency on the total population (live animals) on each time point. We then examined the proportion of carrier and non-carrier animals to obtain the carrier frequencies for each time point. The percentage of affected litters was estimated by taking the product of the carrier frequency, and we examined the piglet loss using the phenotypic records available within the breeding program in the last seven years (2012–2018). To test whether the EL alleles contribute to the heterosis effect for fertility in the crossbred litters in purebred Landrace sows, we made following assumptions: First, we expected no EL litters in the crossbreds (heterozygotes). Second, we assumed 2.9% EL litters in purebred Landrace from the four identified lethal alleles. Third, we calculated the percentage of population deaths for each of the recessive lethals individually by taking the product of affected litters and the litter reduction. Combined, the four lethals account for 0.52% of population deaths, and the overall piglet reduction was calculated as the product of the average TNB (14.17) and the population deaths caused by EL litters in the Landrace population.

6.4.14 Genetic drift simulation

We simulated changes in allele frequency across multiple populations under the model of Wright (Wright 1990). Each simulation was performed with different start frequencies, corresponding to the frequencies of the identified haplotypes. We selected a population Ne of 150, and population size of 2050 (50 boars, and 2000 sows). Each genotype has an associated fitness value, and we set the fitness to zero for homozygous lethal allele carriers, and fitness 1 (no negative fitness effect) to heterozygotes and non-carriers. We assume constant population size through time, and matings are simulated randomly at each generation. Changes in allele frequencies are calculated using the R package driftR (https://github.com/cjbattey/driftR). The simulation calculates allele frequencies from a random draw of a binomial distribution with a probability of success equal to the post-selection expected frequency for each generation and each population. The results are plotted in R using the package ggplot2 (Ginestet 2011).

6.4.15 *De novo* mutations

The frequency of *de novo* mutations was estimated based on a population size of 2050, accounting for a *de novo* mutation allele frequency equal to = 1/4100 = 0.024%. We used a human and cattle based per generation *de novo* mutation rate equal to 1.2^{e-08} per nucleotide per generation (Kong et al. 2012; Harland et al. 2017).

The product of the genome size (in nucleotides) and mutation rate is used to calculate the number of *de novo* mutations per individual (4915.82 Mb * 1.2e-08 = 59). Considering a replacement rate of approximately 50%, we estimate that 60,475 *de novo* mutations will arise each generation (1,025*59). We used the same model from Wright (Wright 1990) to simulate changes in allele frequency across multiple populations for *de novo* mutations.

6.4.16 Breeding values and association analysis

To test whether carriers of lethal haplotypes show signs of heterozygote advantage on important traits in the breeding goal, we performed association analyses between all lethal haplotypes found in this study and a total of 25 traits (S15–S16 Tables) included in the breeding goal of the evaluated populations. Estimated breeding values (EBV) were used as a response variable for each trait under study. The EBV of each animal was obtained from the routine genetic evaluation by Topigs Norsvin using an animal model. Association analyses were performed using the software ASREML (Gilmour et al. 2009) applying the following model:

$$EBVij = \mu + Hi + aj + eij$$

where EBV_{ij} is the observed EBV for the animal j, μ is the overall EBV mean of the population, H_i is the number of copies (0/1) of the lethal haplotype i, a_j is the additive genetic effect estimated using a pedigree-based relationship matrix, and e_{ij} the residual error. A p-value below 1×10^{-5} was considered significant.

6.5 Additional files

The online version of this article (https://doi.org/10.1371/journal.pgen.1008055) contains supplementary material, which is available to authorized users.

6.6 Acknowledgements

The authors wish to thank BioBank AS for collection of samples and DNA extraction, CIGENE at Norwegian University of Life Sciences for providing laboratory facilities. We thank Carole Charlier for her help to identify causal mutations.

7

Detection of a Frameshift Deletion in the *SPTBN4* Gene Leads to Prevention of Severe Myopathy and Postnatal Mortality in Pigs

Martijn F. L. Derks[1], Barbara Harlizius[2], Marcos S. Lopes[2,3], Sylvia W. M. Greijdanus-van der Putten[4], Bert Dibbits[1], Kimberley Laport[1], Hendrik-Jan Megens[1], Martien A. M. Groenen[1]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. [2]Topigs Norsvin Research Center, Beuningen, Netherlands. [3]Topigs Norsvin, Curitiba, Brazil. [4]GD Animal Health Department, Deventer, Netherlands

Frontiers in Genetics (2019) 10:1226.

Abstract

Piglet mortality is a complex phenotype that depends on the environment, selection on piglet health, but also on the interaction between the piglet and sow. However, also monogenic recessive defects contribute to piglet mortality. Selective breeding has decreased overall piglet mortality by improving both mothering abilities and piglet viability. However, variants underlying recessive monogenic defects are usually not well captured within the breeding values, potentially drifting to higher frequency as a result of intense selection or genetic drift. This study describes the identification by whole-genome sequencing of a recessive 16-bp deletion in the *SPTBN4* gene causing postnatal mortality in a pig breeding line. The deletion induces a frameshift and a premature stop codon, producing an impaired and truncated spectrin beta non-erythrocytic 4 protein (SPTBN4). Applying medium density single nucleotide polymorphism (SNP) data available for all breeding animals, a pregnant carrier sow sired by a carrier boar was identified. Of the resulting piglets, two confirmed homozygous piglets suffered from severe myopathy, hind-limb paralysis, and tremors. Histopathological examination showed dispersed degeneration and decrease of cross-striations in the dorsal and hind-limb muscle fibers of the affected piglets. Hence, the affected piglets are unable to walk or drink, usually resulting in death within a few hours after birth. This study demonstrates how growing genomic resources in pig breeding can be applied to identify rare syndromes in breeding populations, that are usually poorly documented and often are not even known to have a genetic basis. The study allows to prevent carrier-by-carrier matings, thereby gradually decreasing the frequency of the detrimental allele and avoiding the birth of affected piglets, improving animal welfare. Finally, these "natural knockouts" increase our understanding of gene function within the mammalian clade, and provide a potential model for human disease.

Key words: animal breeding, loss-of-function, myopathy, pigs, animal welfare

7.1 Introduction

Piglet mortality is one of the major selection traits in pig breeding and is influenced by the sow, the piglets, and the environment. Hence, piglet mortality is a complex phenotype and depends on the capacity of the sow to raise its offspring, but is also a function of birth weight, management, and selection (Knol et al. 2002). However, also monogenic recessive defects contribute to piglet mortality, although only few examples have been reported in the past (Murgiano et al. 2012; Matika et al. 2019). Even in those cases where the effect of the mutation is severe, selecting efficiently against such a mutation is hampered by the low frequency. In many severe defects, zygotes die very early in gestation, leaving no trace other than the absence of homozygotes in the population at large (Derks et al. 2019).

Inbreeding effects in commercial pig populations are usually kept in check by selective breeding for decreased mortality in piglets by improving both mothering abilities and piglet viability (Olijslagers 2018). However, variants underlying recessive monogenic defects are not well captured within the breeding values, and potentially drift to higher frequencies as a result of intense selection (Georges et al. 2019). Moreover, those variants can also be maintained as a result of balancing selection for a correlated positive effect in heterozygous state (Derks et al. 2018).

Recessive defects only marginally contribute the overall piglet mortality (Alonso-Spilsbury et al. 2007). Nevertheless, variants affecting piglet mortality are of great importance because those variants directly influence production and animal welfare (Baxter et al. 2013; Rutherford et al. 2013). However, in animal population management, the low-frequency occurrence of defects is usually poorly documented (often very general terms are used), and syndromes are often only recognized once they have reached a high frequency. This is especially relevant for syndromes that do not lead to very distinct phenotypes. Therefore, even in commercial breeding populations little tracking can be done on specific syndromes, and to effectively select against specific low-frequency syndromes therefore requires new approaches.

In this work, we describe the discovery of a highly debilitating syndrome in a commercial pig population through a survey based on a combined medium-density SNP arrays and whole-genome sequencing (WGS). The survey led to the identification of a 16-bp frameshift deletion in the *SPTBN4* gene, with predicted clear phenotypic consequences in homozygotes. The carrier frequency is about 9% in the population under study, affecting approximately 0.81% of the population litters. The frequency was sufficiently low to be unknown to have a genetic basis, and even effectively being unrecognized as a specific syndrome at all. Upon implementation of the survey, one pregnant sow was identified sired by a carrier boar. The affected piglets suffer from myopathy and are unable to walk, usually resulting in death within a few hours after birth, completely in line with predicted pathology in comparison to similar human and mouse cases.

7.2 Methods

7.2.1 Animals, Genotypes, and Pre-Processing

The dataset consists of 31,839 animals from a synthetic boar line with large white background. The line is maintained and bred in Topigs Norsvin nucleus farms, primarily selecting on production and health traits. The animals were genotyped on the Illumina GeneSeek custom 50K SNP chip (Lincoln, NE, USA). Animals with a frequency of missing genotypes > 0.15 were removed. We discarded markers that did not meet following filtering criteria: A minimum call rate of 0.85, a minor allele frequency > 0.01, and a Hardy-Weinberg proportions exact test p-value below $P < 10^{-12}$. Moreover, markers with unknown location on the Sscrofa11.1 genome build were discarded, leaving 41,573 markers after filtering. All steps were performed in Plink v1.90b3.30 (Purcell et al. 2007).

7.2.2 Haplotype Phasing and Identification of SSC6 Haplotype

We performed haplotype phasing and imputation of missing sites in Beagle5.0 with parameter for effective population size set to 100, other settings were default (Browning et al. 2018). Expected homozygotes was estimated based on haplotype frequency, using the Hardy-Weinberg principle. An exact binomial test was applied to test the number of observed homozygotes with the number of expected homozygotes. The haplotype was considered significantly depleted if $P < 5 \times 10^{-3}$.

7.2.3 Phenotypic Effects Associated With SSC6 Haplotype

We examined the SSC6 haplotype for records on total number born, number stillborn, mummified piglets, farrowing survival, and lactation survival (survival up to about 21 days of age) of a total of 9,666 litters. We listed these phenotypes for all CxC, and CxN litters identified. We used a Welch’s t-test to assess whether the phenotypes from the CxC litters differ significantly from CxN litters. A p-value < 0.05 was considered significant.

7.2.4 Whole-Genome Sequencing Analysis and Candidate Variant Identification

The dataset consists of 71 whole genome sequenced individuals from the population under study. All 71 samples were also present in our dataset of 31,839 animals genotyped on the 50K. The 71 samples have a total volume of 1.93 Tbp (tera base pairs) from 14.16 billion 150-bp paired-end reads (Table S3). The samples were sequenced on Illumina HiSeq 2000. We aligned the sequences to the Sscrofa11.1 genome build using BWA-MEM version 0.7.15 (Li and Durbin 2009) with an average mappability of 98.9% and a sample coverage ranging from 8.8 to 14.8X (10.9X average). Samblaster was used to remove PCR duplicates (Faust and Hall 2014). Samtools was used to sort, merge, and index bam files (Li et al. 2009). Mapping and quality statistics were generated using Qualimap (Okonechnikov et al. 2016). Variant calling was performed with Freebayes v1.1.0 with following settings: –min-base-quality 10 –min-alternate-fraction 0.2 –haplotype-length 0 –min-alternate-count 2 (Garrison and Marth 2012). Variants with Phred quality score < 20 were discarded (Li et al. 2009). Variants were annotated using the Ensembl variant effect predictor (VEP, release 96) (McKenna et al. 2010b). The impact of missense variants was predicted using sorting intolerant from tolerant (SIFT) (Kumar et al. 2009). LD analysis was performed using Plink v1.90b3.30 (Purcell et al. 2007) with following settings –chr-set 18, –r2, ld-window-r2 0.8.

7.2.5 SPTBN4 Protein Alignment

Protein alignment between the wild type and mutant protein was performed using ClustalO (Madeira et al. 2019) and visualized using ESPript 3 (Robert and Gouet 2014). Further visualization and validation was performed using the JBrowse genome viewer version 1.12.1 (Skinner et al. 2009).

7.2.6 Validation of Causal 16 bp *SPTBN4* Deletion

PCR was done using 60 ng of genomic DNA, with 0.4 µm of each primer, 1.8 mM MgCl2, and 25 units/ml OneTaq® DNA Polymerase (OneTaq® 2X Master Mix with Standard Buffer, New England Biolabs) in manufacturer's PCR buffer in a final volume of 12 µl. Initial denaturation for 1 min at 95°C was followed by 35 cycles of 95°C for 30 s, 55°C for 45 s, 72°C 90 s, followed by a 5 min extension 72°C. PCR primers for *SPTBN4* are TCAAGGGTGCAGGCTCTTTC forward and GGTAGGAAGCTCGAAGTGGG reverse. The forward primer was dye-labeled with either 6-FAM to produce a fluorescently labeled PCR product detectable on ABI 3730

DNA sequencer (Applied Biosystems). Fragment sizes were determined using GeneMapper software 5 from ABI.

7.2.7 Histopathological Examination

Two affected piglets less than 1 week old were send to the pathology department of Royal Animal Health (Deventer) for examination. Macroscopically, all observations were within normal limits. Skeletal muscle of the foreleg, the dorsal muscle, and the backside leg of both animals was sampled for routine H&E staining and PTAH staining. The muscle tissue was stored in separate jars and fixated in formaldehyde solution 4%, buffered (=formalin solution 10%, buffered). After that, the tissue was embedded in paraffin and sliced into 2 µm according to standard operation procedure (SOP RAH). Thereafter, the slides were deparaffinized and routinely stained for hematoxylin and eosin (H&E) in an automatic color machine. Simultaneously additional slides of 2 µm of the muscle tissue as well as a positive control slide of muscle tissue were prepared for the manual staining with "phosphotungstic acid hematoxylin," abbreviated as PTAH. This staining is preferred for demonstrating cross-striations of skeletal muscle.

7.2.8 Breeding Values and Association Analysis

In this study, we evaluated 63 traits used in the breeding program. Deregressed estimated breeding values (DEBV) were used as a response variable for each trait under study. The estimated breeding value (EBV) of all evaluated traits were deregressed using the methodology described by (Garrick et al. 2009). The EBV of each animal was obtained from the routine genetic evaluation by a commercial breeding program (Topigs Norsvin) using an animal model. The reliabilities per animal for the purpose of deregression were extracted from the genetic evaluation based on the methodology of (Tier and Meyer 2004). The heritabilities used for the deregression were also extracted from the routine genetic evaluation. Finally, weighting factors based on the estimated reliability of the DEBV were also estimated according to Garrick et al. (2009) using a value of 0.5 for the scalar c. To ensure the quality of the DEBV, only animals with a weighting factors greater than zero and a reliability of the DEBV greater than 0.20 were used in the association analyses. The reliability of the DEBV was also obtained according to Garrick et al. (2009).

Association analyses were performed using the software ASREML (Gilmour et al. 2009) applying the following linear mixed animal model:

$$DEBVij\omega \;=\; \mu + Ri + aj + eij,$$

where DEBV$_{ij}$ is the observed DEBV for the animal *j*, w is weighting factor for the residual, μ is the overall DEBV mean of the population, R_i is the carrier status (count of the detrimental allele) of the *4* mutation *i*, a_j is the additive genetic effect estimated using a pedigree-based average relationship matrix, andthe residual error. Associations with a −log10(P value) greater than five were declared as significant.

7.3 Results

7.3.1 A 1.5 Mb Segment on Chromosome 6 Affects Lactation Survival in Pigs

We analyzed 31,638 animals from a single purebred boar line (synthetic line with large white background), genotyped on the Porcine 50K SNP chip (Sscrofa11.1 build) (Warr et al. 2019). The analysis revealed a 1.5 Mb segment on chromosome 6 (SSC6:48.75–50.25) showing a deficit in homozygosity associated with reduced lactation survival (Tables 7.1 and 7.2). The haplotype is segregating at a moderate allele frequency of 4.5% (9.0% carrier frequency) in the population under study. The haplotype frequency has been fluctuating over the last decade, but decreased over the last 3 years (Figure S1). We tested whether the frequency was driven by an heterozygous advantage effect. However, we found mostly negative associations with important selection traits except for loin depth and gestation length (Table 7.3), which suggests the frequency is purely the result of genetic drift.

Table 7.1 SSC6 haplotype characteristics. Shown are the expected and observed homozygotes for the SSC6 haplotype.

Position, Mb	SSC6: 48.75–50.25
Number of markers	19
Homozygotes expected (HWE)	61.07
Homozygotes observed	0
Exact binomial test	3.04e−27
Carrier frequency %	9.0
C x C matings	52
Genotyped C x C progeny	73
Heterozygote C x C progeny	46 (63.0%)

The 52 carrier-by-carrier (CxC) litters show no significant reduction in total number born or liveborn animals. However, lactation survival is reduced by about 24% in CxC litters compared to carrier-by-noncarrier (CxN) matings, indicating that homozygous piglets die within the lactation period (Table 7.2). Next, we examined the remarks for time and cause of mortality of CxC litters. This revealed that most piglets that died within the first 24 h after birth. The majority of those piglets were mostly described by farmers as "weak piglet at birth."

Table 7.2 Carrier-by-carrier litters show 24% decrease in lactation survival compared to carrier-by non-carrier litters. Significant results are indicated in bold. The sow is carrier in CxN litters, while the boar is carrier for NxC litters.

Status	# Litters	Avg. total born	Avg. live-born	Farrowing survival %	Lactation survival %
NxN	8,105	10.08	9.23	91.37	89.97
CxN	732	10.44	9.62	92.09	90.80
NxC	777	10.16	9.43	92.78	89.85
CxC	52	9.96	9.13	91.82	68.84*

* P < 0.01

Table 7.3 Traits significantly associated with heterozygous carriers of the *SPTBN4* deletion. Effect shows the direction of the association, SE shows the standard error. The symbols "+" and "−" indicate positive and negative effects.

Trait	Non-carriers	*SPTBN4* carriers	Effect	SE	$-\log_{10}$(P value)
Lactation (pre-weaning) survival$^-$	13,789	1,506	−0.39	0.04	20.84
Intramuscular fat (loin)$^-$	5,676	696	−0.08	0.01	17.66
Lifetime daily gain$^-$	14,089	1,548	−5.62	0.7	14.77
Daily feed intake$^-$	14,081	1,548	−16.18	2.99	7.19
Loin depth at end of test period$^+$	14,089	1,548	0.28	0.06	6.45
Litter mortality $^-$	13,442	1,462	0.21	0.05	5.25
Gestation length$^+$	14,051	1,544	−0.08	0.02	5.04

7.3.2 Whole-Genome Sequencing Analysis Reveals a 16-bp Frameshift Deletion in *SPTBN4* as the Likely Causative Variant

To identify the causal mutation, we examined whole-genome sequence data from 71 animals from the population under study and identified five carrier animals. Linkage disequilibrium (LD) analysis revealed 267 SNP and indel variants in high LD ($r^2 > 0.8$) with the SSC6 haplotype (Table S1), the majority being in perfect LD (247 variants). Only five variants potentially affect the coding sequence (three missense, one frameshift, one splice-acceptor). The three missense variants are predicted to be tolerated by SIFT (score > 0.18, Table S1), while the splice-acceptor variant affects a gene encoding a 28 bp peptide of unknown function, unlikely to be causal. However, one variant in complete LD ($r^{2\,=}$ 1) with the haplotype was predicted to have high impact; a 16-bp frameshift deletion in exon 26 of the *SPTBN4* gene (6:g.48801280delGACGGTGTACGCCGGT) (Figures 7.1A, 7.1B). The frameshift deletion (ENSSSCP00000031537:p.Arg1902fs) introduces 30 novel amino acids and a premature stop codon, producing an impaired and truncated spectrin beta non-erythrocytic 4 protein (SPTBN4). Mutants lack the final 662 amino acids of the wild type protein (Figure 7.1C), including the pleckstrin homology (PH) domain required for protein transport to membranes . The SPTBN4 protein is a member of the beta-spectrin proteins and is (Wang et al. 2018)an actin that links the cell membrane to the actin cytoskeleton. *SPTBN4* mutations disrupt the cytoskeletal machinery controlling proper localization of ion channels in myelinated nerves causing motor neuropathies (Parkinson et al. 2001; Wang et al. 2018).

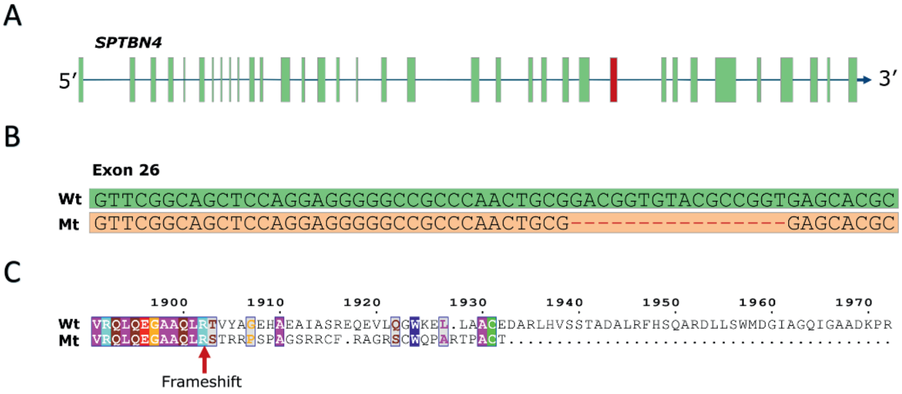

A
SPTBN4
5′
3′
B
Exon 26
Wt GTTCGGCAGCTCCAGGAGGGGGCCGCCCAACTGCGGACGGTGTACGCCGGTGAGCACGC
Mt GTTCGGCAGCTCCAGGAGGGGGCCGCCCAACTGCG----------------GAGCACGC
C
1900 1910 1920 1930 1940 1950 1960 1970
Wt VRQLQEGAAQLRTVYAGEHAEAIASREQEVLQGWKEL.LAACEDARLHVSSTADALRFHSQARDLLSWMDGIAGQIGAADKPR
Mt VRQLQEGAAQLRSTRRPSPAGSRRCF.RAGRSCWQPARTPACT..
Frameshift

Figure 7.1 (A) SPTBN4 gene model. The location of the affected 26th exon is indicated in red. **(B)** Illustration of the 16-bp deletion. Figure shows wild type and mutant exon. **(C)** Alignment of the mutant (Mt) and wild type (Wt) SPTBN4 protein sequence. The mutation induces 30 novel amino acids and a premature stop codon.

7.3.3 Genotyping Five CxC Litters Confirms *SPTBN4* Deletion as the Likely Culprit

We genotyped five CxC litters for the 16-bp deletion which had at least two piglets (range 2–6) that died within the first 48 h after birth. The five litters produced 53 piglets of which 19 were homozygous for the 16 bp deletion (Table 7.4). All 19 homozygous piglets died within 48 h after birth (18 within 24 h). From the 34 remaining piglets (8 wild type, and 26 carriers), only 1 died within 48 h, likely caused by other (environmental) factors.

Table 7.4 Genotyping of the likely causal 16-bp *SPTBN4* frameshift deletion in five carrier-by-carrier litters. The sum per genotype class is indicated in bold. All animals homozygous for the deletion died within 48 h after birth.

Litter ID	# Genotyped progeny	# Wild type	# Deletion carrier	# Homozygotes
1	12	3	5	4
2	12	0	8	4
3	11	0	5	6
4	8	1	5	2
5	10	4	3*	3
SUM	**53**	**8 (15.1%)**	**26 (49.1%)**	**19 (35.8%)**

*One piglet died within 24 h after birth.

7.3.4 Piglets Homozygous for the *SPTBN4* Deletion Suffer From Myopathy and Hind Limb Paralysis

We monitored one recent CxC litter (farrowing date: April 28th2019) that produced six healthy, two affected (samples: 9912, 9916) (Figure 7.2A), and three stillborn piglets. We confirmed the homozygous *SPTBN4* deletion status for the two affected piglets (Table S2). Moreover, we observed four heterozygous carriers and two homozygous wild type piglets among the healthy individuals. One of the stillborn piglets (sample: 9921) was also homozygous for the deletion, while the other two were heterozygous. The affected piglets suffer from extreme muscle weakness (Figures 7.2B, 7.2C), paralysis of the hind limbs, and tremors (S1 Video). Hence, the piglets were unable to walk or drink.

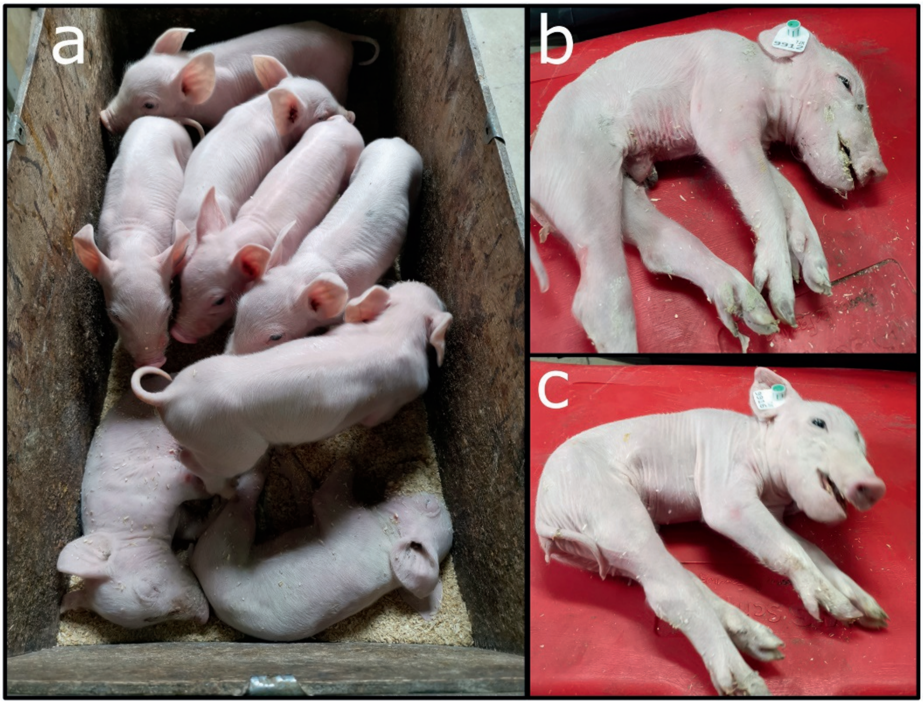


Figure 7.2 (A) Two affected piglets (alive) together with six healthy littermates. The piglets derive from one CxC mating farrowed on 28th of April 2019. **(B)** Affected male piglet 9912. **(C)** Affected female piglet 9916.

7.3.5 Affected Piglets Lack Cross Striations in the Dorsal and Hind Limb Skeletal Muscles

Histopathological examination revealed scattered degeneration of muscle fibers in both piglets, and focally necrosis and vasculitis in the dorsal muscle in one of the piglets (ID = 9912). Moreover, phosphotungstic acid hematoxylin (PTAH) staining shows divergent coloring of the skeletal muscle fibers, indicating decrease of cross-striations, particularly in the muscles of the dorsal and hind legs of the affected animals (Figure 3B), while the front legs seem unaffected (Figure 7.3A). The decrease of cross striations is indicated by abnormal coloring and general loss in volume of muscle fibers (Figure 7.3B). The histopathologically observed changes in the hind legs and in the dorsal muscles are indicative for muscular dystrophy.

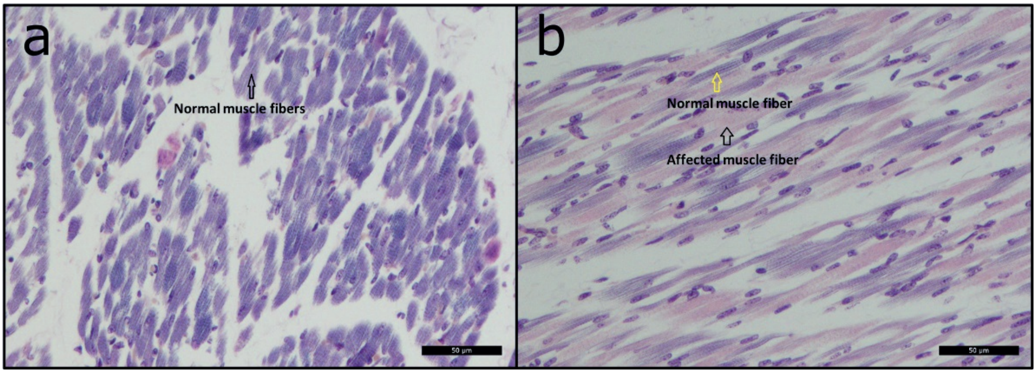


Figure 7.3 (A) Cross-sectional view of a skeletal muscle from the front leg. The black arrow indicates normal coloring (dark) of muscle fibers indicating presence of cross-striations. PTAH Bar = 50 µm. **(B)** Cross-sectional view of a skeletal muscle from the hind leg. The black arrow indicates abnormal coloring (pink) of muscle fibers indicating lack of cross-striations. The yellow arrow indicates normal coloring and presence of cross-striations. PTAH Bar = 50 µm.

7.4 Discussion

In this work we report a novel congenital defect causing piglet mortality likely due to a 16 bp frameshift deletion in the *SPTBN4* gene. The piglets suffer from extreme muscle weakness (myopathy) and die within a few hours after birth. The deletion is expected to confer a complete loss-of-function of the spectrin beta, non-erythrocytic 4 protein. *SPTBN4* is a member of the family of spectrin genes and is required for ion channel clustering at the nodes of Ranvier, affecting action potential (Devaux 2010). Mutations disrupt the cytoskeletal machinery that controls proper localization of ion channels and function of axonal domains mainly at the axon initial segments (AIS) and the nodes of Ranvier (Wang et al. 2018). More specifically, the affected C-terminal domain of SPTBN4 is crucial for KCNQ2 channel trafficking and excitability at nodes of Ranvier (Devaux 2010).

Subsequent follow-up research identified human and mouse cases that indicated that the ensuing syndrome would likely not prove to be immediately lethal, but rather confer severe myopathy. By medium-density SNP genotype data, available for all animals in the breeding population (N = 31,839), carriers could be identified. Among those carriers was a sow that was approximately mid-term in pregnancy at the time of identification, sired by a boar that was also carrier. The breeding farm was notified to document the litter at birth. The observed phenotype of the affected piglets (myopathy, hind limb paralysis, tremors) was completely congruent with what was observed in human patients with homozygous loss-of-function or compound heterozygous mutations in the *SPTBN4* gene (OMIM: 606214). Two of the human patients have loss-of-function mutations within the PH domain (Wang et al. 2018), supporting that a loss of the PH domain in pigs would likely lead to a complete loss-of-function of the SPTBN4 protein. In human, similar mutations lead to severe congenital myopathy caused by the absence of muscle type I fibers, neuropathy, and deafness (Knierim et al. 2017; Wang et al. 2018). Wang et al. (2018) also observed motor axonal neuropathy in several patients characterized by congenital hypotonia, profound weakness, and loss of deep tendon reflexes by early childhood. Moreover, nerve biopsies revealed reduced nodal Na+ channels and no nodal KCNQ2 K+ channels, revealing the molecular pathology causing nervous-system dysfunction. Therefore, we conclude that this frameshift variant is the likely causal mutation leading to the observed phenotype and depletion of the homozygous genotype in the population. Future studies could focus on making an *in vivo* knockout of the *SPTBN4*gene in pig, to study the syndrome and associated phenotype in more detail.

We did not observe degeneration of muscle fibers in the front legs, while the dorsal and hind leg muscle fibers were clearly affected. This observation could partly explain the hind limb paralysis, while the front legs are not affected. The discrepancy between front and hind legs muscle fibers has also been described in quivering mice, in which *SPTBN4* loss-off-function mutations cause motor neuropathy, hind limb paralysis, tremors, and central deafness (Parkinson et al. 2001; Komada and Soriano 2002). Parkinson et al. (2001) describe reduced nerve-conduction velocities in sciatic nerves of mice with quivering alleles causing the peripheral hind limb neuropathy. Expression of *SPTBN4* in mice is restricted to the brain, spinal cord, and sciatic nerves and not observed in skeletal muscle, so this disease is primarily a neuronal defect. Overall it remains unclear which mechanism causes the absence of symptoms in the forelimbs. This “natural knockout” in pigs can be a useful resource to study the human disease, as pigs are usually a better model to study human disease compared to rodent species. Moreover, the consequence of the loss of *SPTBN4* function can be studied in more detail.

The effective population size (Ne) of the breed under study is estimated to be around 100 (Hidalgo et al. 2016). In animal breeding, low Ne increases the risk that detrimental alleles rise in frequency by chance. Moreover, previous studies have shown that recessive lethal alleles can be driven by advantageous effects in heterozygotes (Derks et al. 2018; Matika et al. 2019). Matika et al., 2019 found a recessive stop-gained mutation in the *MSTN* gene associated with a major increase in muscle depth in heterozygotes. However, we find no evidence for any heterozygous advantage in our study. With the current genomic techniques we can now identify deleterious alleles drifting to higher frequencies, and monitor the emergence of novel deleterious alleles accurately, allowing more effective purging. Moreover, the result of this type of study will greatly improve the consciousness of "hidden" genetic defects at both the breeder and farmer level. Without any prior information, rare birth defects are often recorded as "weak piglet." And without any further distinction of specific syndromes, further action is not possible. In most cases it is unknown if there is a genetic basis, or that there may be other confounding effects. With prior genomic information, the syndrome can be identified, compared to other cases, and carriers identified, leading to actionable information.

Piglet mortality is of high economic and animal welfare importance. Hence, the discovery of the *SPTBN4* mutation has led to immediate implementation in the breeding program to minimize the frequency of carrier-by-carrier matings. This enables to avoid the birth of affected individuals, thereby improving animal welfare and reducing economic losses.

7.5 Conclusion

In this study we report a novel congenital defect likely caused by a recessive frameshift deletion in the *SPTBN4* gene in pigs. The findings are supported by striking similarities to *SPTBN4* associated syndromic phenotypes in humans and mice. The study allows to monitor and purge the deleterious allele from the population. Carrier-by-carrier crosses can be prevented, precluding affected individuals, thereby reducing economic losses, and improving animal welfare. Finally, these "natural knockouts" obtained in the breeding industry can provide a model for human disease and increase our understanding of gene function within the mammalian clade, and provide a potential model for human disease.

7.6 Ethics statement

Ethical review and approval was not required for the animal study because the data used in this study has been obtained as part of routine data collection from Topigs Norsvin breeding programs, and not specifically for the purpose of this project. Therefore, approval of an ethics committee was not mandatory. Sample collection and data recording were conducted strictly according to the Dutch law on animal protection and welfare (Gezondheids- en welzijnswet voor dieren). Written informed consent was obtained from the owners for the participation of their animals in this study.

7.7 Additional files

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01226/full#supplementary-material

7.8 Acknowledgements

The authors thank Frank van Haaren and Toon Janssen for organising communication with the farm. We thank Manon Houben for her contribution to the pathological examinations. We thank Mout Viller for taking video and pictures of the affected and healthy piglets after birth. We thank Jürgen Harlizius for taking care of the piglets after handover. We thank Gerda van Eldik for sample collection. We thank Egbert Knol and Egiel Hanenberg for general support and guidance from Topigs-Norsvin.

7.9 Authors' contributions

MG, H-JM, and MD conceived and designed the study. BH was responsible for general organisation and communication with Topigs Norsvin and farmers. MD and ML performed the data analysis. BD and KL performed lab work. SG-V performed the pathological analysis. MD wrote the manuscript. H-JM, MG, BH, SG-V, BD, KL, and ML provided useful comments and suggestions and helped to draft the manuscript. Phenotypic data was analysed by ML. All authors read and approved the final manuscript.

8

Accelerated discovery of functional genomic variation in pigs

Martijn F.L. Derks[1,*], Christian Gross[2,3,*], Marcos S. Lopes[4,5], Marcel J.T. Reinders[3], Mirte Bosse[1], Arne B. Gjuvsland[6], Dick de Ridder[2], Hendrik-Jan Megens[1], Martien A.M. Groenen[1]

1 Wageningen University & Research, Animal Breeding and Genomics, Wageningen, The Netherlands. 2 Bioinformatics Group, Wageningen University and Research, P.O. Box 633, 6708 PB, Wageningen, The Netherlands. 3 Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands. 4 Topigs Norsvin Research Center, Beuningen, the Netherlands. 5 Topigs Norsvin, Curitiba, Brazil. 6 Norsvin SA, Hamar, Norway

In preparation

Abstract

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

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

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

8.1 Introduction

Closing the gap between genotype and phenotype is a major goal in many life sciences, but remains extremely challenging (Gjuvsland et al. 2013). Part of the complexity arises from the polygenicity of phenotypes, in which many (interacting) loci contribute to the observed phenotype. Genome wide association studies (GWAS) have been instrumental to associate genomic markers to important phenotypes reported as quantitative trait loci (QTL), and to get a better grip on the biology of the traits (Schaid et al. 2018). However, the resolution of GWAS is limited by the correlation between neighbouring markers in linkage disequilibrium (LD). Hence, unravelling the molecular drivers underlying phenotypes of interest requires the identification of the actual causal variants (Gallagher and Chen-Plotkin 2018), which often reside in the noncoding regions of the genome, in particular in predicted transcriptional regulatory regions (Ponting and Hardison 2011).

In human genetics, a combination of statistical fine-mapping methods and expression QTL (eQTL) studies are used to further narrow down the list of candidate causal variants (Cannon and Mohlke 2018). Further functional annotation, facilitated by large consortium efforts like the Encyclopedia of DNA Elements (ENCODE, (ENCODE 2012)), is used to prioritize variants based on likelihood of affecting a regulatory region, affecting gene expression. Despite this effort, identifying the causal variant remains difficult, partly because of the fundamental complexity of phenotype-genotype relations, in which also the environment plays an important role.

Also in livestock, economically important phenotypes are typically determined by a very large set of variants each explaining a small fraction of the phenotypic variation. However, for many trait there are also some QTLs explaining a larger fraction (>1%) of the variation. For such larger QTLs it is of interest to identify the underlying causal variation. Due to intense selection, the effective population size (N_e) of most livestock populations is small (Hall 2016). This often leads to extended LD, comprising up to millions of basepairs (Mb) in length, especially in regions with low recombination rates (Veroneze et al. 2013). High LD yields an additional layer of complexity to fine-map GWAS results in livestock populations, and the use of crossbreeding to break down the LD is a costly, labor-intensive and time-consuming procedure to fine map the QTL region. On the contrary, livestock populations are less confounded by population stratification (i.e. ancestry differences between cases and controls), which can be a major factor in human GWAS studies (Hellwege et al. 2017).

Similar to human, further functional genomic information could help to prioritize the variants underlying the phenotypes of interest in livestock (Ron and Weller 2007). However, in pigs, the level of functional genomics information is limited. Fortunately, recent advances have been achieved in pigs by the publication of the pig Combined Annotation-Dependent Depletion (pCADD) tool (Gross et al. 2019), providing impact scores of any nucleotide substitution in the pig genome. CADD was developed to score variants with respect to their putative deleteriousness to prioritize potentially causal variants in genetic studies (Rentzsch et al. 2019). This tool is frequently used to score variants in human GWAS studies (Cannon and Mohlke 2018). Subsequently, other species-specific CADD tools were developed (Gross et al. 2018). The tool scores the deleteriousness (or functional impact) of single nucleotide variants (SNPs), and is built on many layers of annotations including sequence context, conservation scores, gene expression data, non-synonymous mutation scores, and epigenomic data, if available for the investigated species.

Pig populations have been under a long-term biological experiment by animal breeders that use genomic selection to constantly improve their stock (Knol et al. 2016). In general, genomic selection uses a variant panel on a chip to associate regions in the genome with important traits. This variant panel is distributed across the genome and allows within-population genetic variation to be captured (Meuwissen et al. 2001). However, genomic selection uses the genome as a “black box”, as the SNPs on the chip are mostly not causal, but genetically linked to the actual causal variants and genes (Habier et al. 2013). Therefore, the efficacy of genomic selection can be substantially improved by adding new genetic markers comprising the actual causal variation (Goddard et al. 2016), providing insight in the exact molecular drivers involved in the selection.

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

8.2 Results

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

We analysed large-scale genotype and phenotype data in four purebred pig populations: two boar breeds of Duroc and Synthetic origin, and two sow breeds of Landrace and Large White origin. In pigs, selection takes place on the purebred populations, while the final production animals are derived from three-way crosses. First, crossbred sows are created from populations selected for high reproductivity and mothering abilities, which are subsequently crossed with a population especially selected for meat production traits. The examined traits can be grouped in three classes: (1) traits focusing on carcass and meat quality, including backfat, intramuscular fat, and growth; (2) reproduction traits, mainly focusing on litter size, number of liveborn, survival, and mothering abilities; and (3) health and welfare traits including disease resistance, osteochondrosis, umbilical hernia, and other conformation traits. A total of 129,336 animals with 552,000 imputed SNPs were subjected to a GWAS analysis for 83 traits. The analysis revealed a large set of QTL regions with a genome-wide association significance threshold of -log10(p)>6.0, and significant associations were observed for the majority of examined traits. The 'lead' SNP that showed the strongest association signal is used as a starting point for further analysis.

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

8.2.2.1 pCADD scores all possible substitutions from the Sscrofa11.1 pig reference genome.

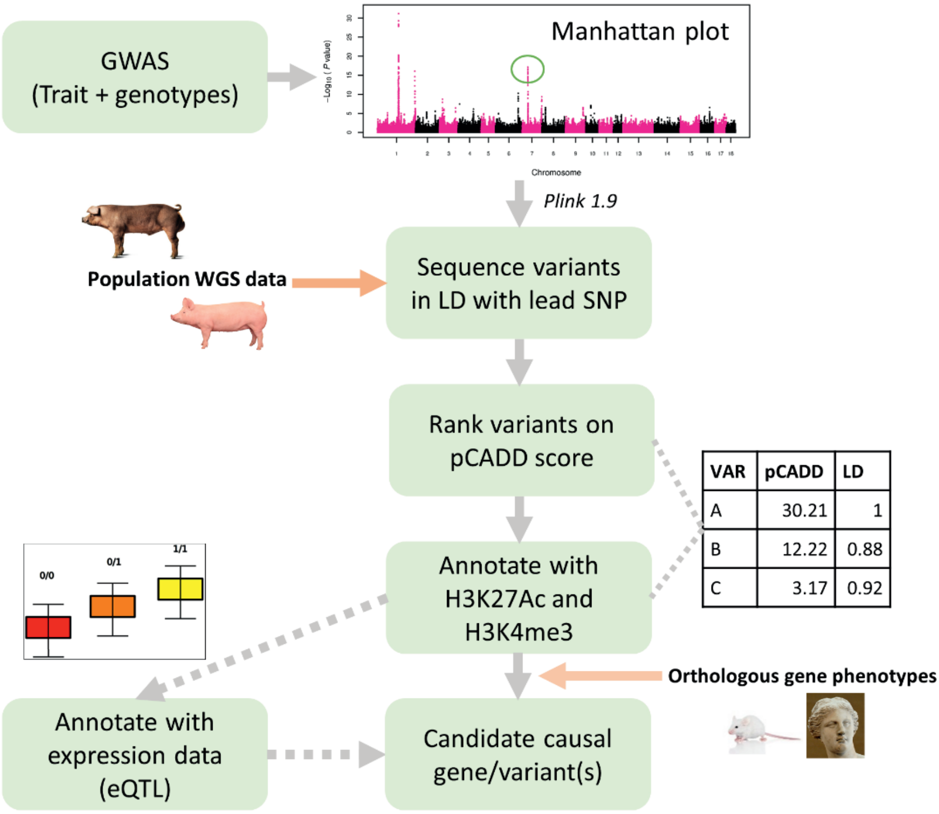
Our approach first relies on the lead SNP from a significant GWAS peak to extract sequence variants that are in high LD (r^2>0.7). The whole-genome sequence variants are extracted from a total of 428 animals (Duroc: 101, Synthetic: 71, Landrace: 167, Large White: 89), sequenced to an average depth of 11.82. Next, we assigned pCADD scores to each sequence variant in high-LD with the lead SNP, to prioritize them on their likely impact. The sequence variants were assigned to a functional class using the Ensembl Variant Effect Predictor (VEP, release 98) (McLaren et al. 2016). The distribution of the pCADD scores for a set of variants depends on their functional class, and non-coding variants have on average lower scores compared to coding variants. The quantiles and further class statistics for the pCADD scores are presented in Table S1. In addition, three liver histone modification datasets were used (for modifications H3K27Ac and H3K4me3) to mark variation overlapping with regulatory sequences, including likely active promoter and enhancer elements in pig liver tissue (Villar et al. 2015).

8.2.2.2 Phenotype and pathway information provides further evidence of gene causality

Functional annotations, including pathways and gene-ontology information for the examined pig genes associated with the top-ranked variants were extracted from the Uniprot database (UniProt 2019). Moreover, we extracted associated phenotypes from orthologous genes from the Ensembl database for human (*Homo sapiens*), mouse (*Mus musculus*), and rat (*Rattus norvegicus*). The phenotypes are mainly based on (disease) association studies in human, and gene-knockouts in mouse and rat (Zerbino et al. 2018). A complete overview of the pipeline is presented in Figure 8.1.

8.2.2.3 Gene expression information allows identification of possible expression quantitative trait loci

The combination of genotype and gene expression data provides an additional layer of evidence to find causal variation, as differences in expression of genes can be associated with a variant (expression quantitative trait loci; eQTL). In this study we use 59 RNA-sequenced samples (van Son et al. 2017b) from Landrace (n=34) and Duroc (n=25) to test for differential expression between the genotype classes (homozygous reference, heterozygous, homozygous alternative) to associate the expression of genes with the genotypes. The samples were sequenced from testis tissue and further details about the sequenced samples and alignment depth are provided in Table S2. The combination of epigenomic marks (liver) and gene-expression data (testis) can, on top of the pCADD scores, facilitate in the discovery of functional variants.


GWAS
(Trait + genotypes)
Manhattan plot
−Log10 (P value)
Chromosome
Plink 1.9
Population WGS data
Sequence variants in LD with lead SNP
Rank variants on pCADD score
0/0
0/1
1/1
Annotate with H3K27Ac and H3K4me3
Orthologous gene phenotypes
Annotate with expression data (eQTL)
Candidate causal gene/variant(s)

VAR	pCADD	LD
A	30.21	1
B	12.22	0.88
C	3.17	0.92

Figure 8.1 Pipeline overview. The pipeline takes the result of a GWAS as input (lead SNP) and identifies SNPs from WGS data that are in high LD with the lead SNP. Subsequently, the variants are prioritizes based on impact scores (pCADD), open chromatin information (liver), and gene expression (if available). The pipeline outputs a final list of candidate causal variants for each trait of interest, ranked on its likely importance.

8.2.3 Accelerated discovery of potential causal variants from GWAS results

To demonstrate the utility of our approach we first analysed several QTL regions with known causal variants reported in literature. This list includes a missense mutation in *MC4R* affecting production traits (Kim et al. 2000), a promoter variant affecting number of teats in the *VRTN* gene (van Son et al. 2019), and a missense mutation affecting meat quality in *PRKAG3* (Milan et al. 2000). The method returned the causal variant as top ranked for both the *MC4R* missense mutation (Text S1, Figure S1) and the *VRTN* promoter variant (Text S2, Figure S2, Table S4), despite the fact that hundreds of variants were found in LD with the lead SNP.

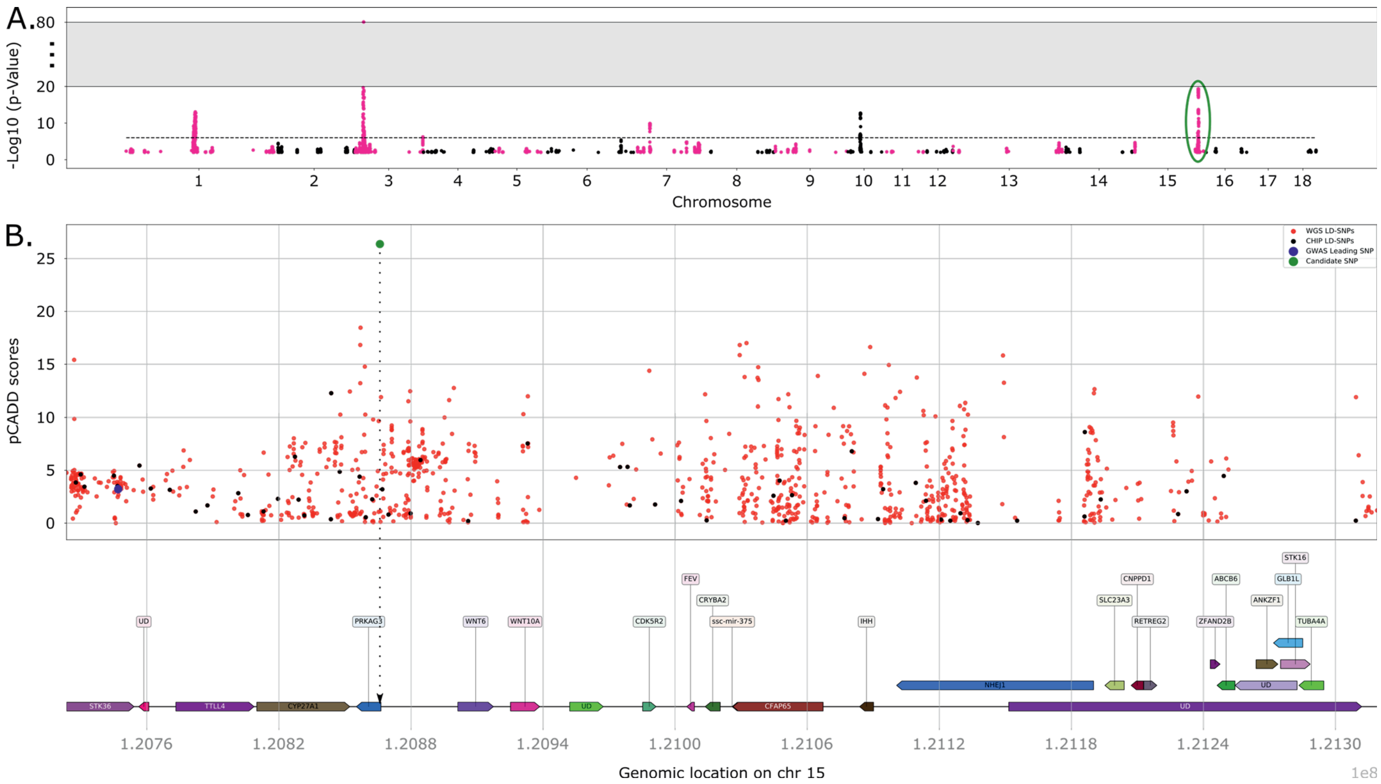


Figure 8.2 A) Manhattan plot for drip loss in Duroc showing a strong QTL on chromosome 15:121Mb. Only SNPs with a -log$_{10}$(p) > 2 are plotted. B) Plot showing all sequence variants in high LD (red) with the lead SNP (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98.

The mutation identified by Milen et al (2000) does not segregate in our sequenced animals, however, we identified another missense variant (15:g.120865869C>T) in the *PRKAG3* gene likely affecting meat quality in both boar breeds (Figure 8.2), as described by UImari et al. 2014 (Uimari and Sironen 2014). The causal missense variant is highlighted in green, and the lead SNP in the GWAS results in blue in Figure 8.2B. The variant substitutes glutamic acid for lysine (ENSSSCP00000030896: p.Glu47Lys) and is segregating at a frequency of approximately 20%, and 36% in Synthetic and Duroc, respectively (Table S5). *PRKAG3* regulates several intracellular pathways, including glycogen storage (Essen-Gustavsson et al. 2011). The specific isoform (ENSSSCT00000036402.2) affected by the Glu47Lys missense mutation has a role in the metabolic plasticity of fast-glycolytic muscle and is primarily expressed in white skeletal muscle fibers (Mahlapuu et al. 2004). Gain of function mutations in the *PRKAG3* gene have been correlated with increased glycogen content in skeletal muscle in pig, negatively affecting meat quality (Ciobanu et al. 2001). The Lys47 variant likely causes a gain-of-function of the 5'-AMP-activated protein kinase subunit gamma-3 enzyme, resulting in increased glycogen content causing lower water holding capacity resulting in low meat quality.

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

8.2.4.1 Promoter variants in the HMGA1 and HMGA2 genes affects fat deposition and growth in pigs.

A strong QTL on chromosome 7 affects backfat, intramuscular fat, growth, feed intake and loin depth in Duroc (Figure 8.3A). The lead SNP in the GWAS result is located at position 7:30,116,227 with a $-\log_{10}(p) > 20$ for backfat, feed consumption, and intramuscular fat (Figure S4). The analysis returns 485 variants in high LD with the lead SNP (Figure 8.3B). The two variants with highest pCADD scores are annotated upstream of the *HMGA1* gene, 566 bp apart (Figure 8.3B, Table S6). Both mutations are in the promoter region of the *HMGA1* gene, supported by signals on the H3K4me3 and H3K27Ac histone marks (Figure S5). The A allele, segregating at 36% allele frequency, is associated with less backfat, faster growth, but also smaller loin and decreased intramuscular fat. We evaluated the expression of the *HMGA1* gene in twenty samples for which both genotype and gene expression, as normalized fragments per kilobase per million (FPKM), were available within the three genotype classes GG, AG, and AA. The A allele causes increased expression of the gene in an additive manner (P=0.041, Figure S6) and suggests that increased expression of the *HMGA1* gene positively affects backfat and growth, but decreases intramuscular fat.

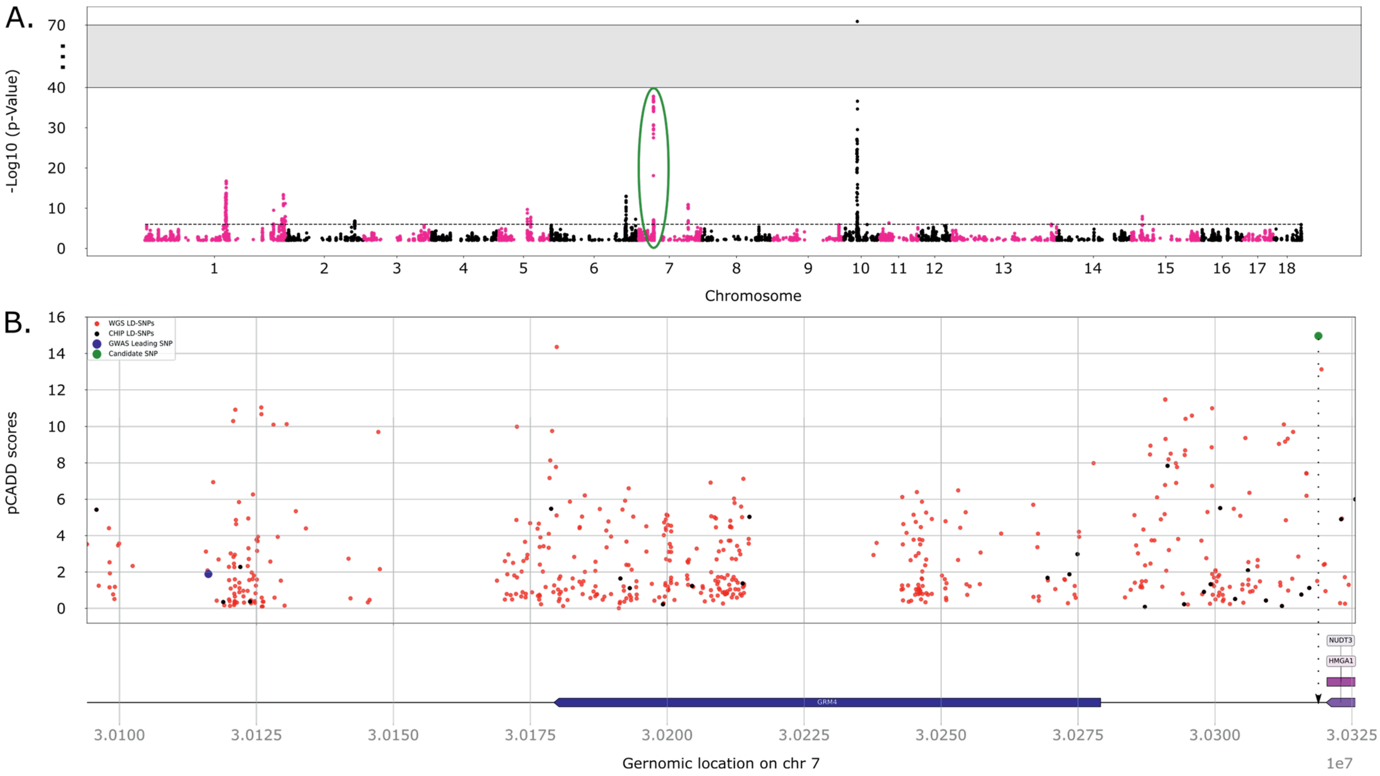


Figure 8.3 A) Manhattan plot for backfat in Duroc showing a strong QTL on chromosome 7:30Mb. Only SNPs with a -log$_{10}$(p) > 2 are plotted. B) Plot showing all sequence variants in high LD (red) with the lead SNP (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98

In addition, we find two variants affecting the promoter region of the *HMGA2* gene, to be associated with less backfat in the Synthetic breed (Table 8.1, Table S7). Both *HMGA1* and *HMGA2*, part of the High Mobility Group A gene family, are well-known genes to affect growth and stature in pigs (Kim et al. 2006; Hong et al. 2015; Chung et al. 2018), but no causal variant has been reported thus far. Our results suggest that the causal variants for both genes are regulatory variants.

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

A strong QTL on chromosome 1 affects backfat, intramuscular fat, and drip loss in the Synthetic breed (Figure 8.4A). The lead SNP in the GWAS result is located at position 1:115,884,118. The analysis returns 874 variants in high LD with the lead SNP (Table S8). The SNP with the highest pCADD score (1:g.120074006G>A), a single missense variant affecting the *SCG3* gene is identified as the likely culprit (Figure 8.4B). The variant substitutes a threonine for a methionine at position 386 in the Secretogranin-III protein (ENSSSCP00000044507:p.Met386Thr). The Met386 allele is associated with increased intramuscular fat, more backfat and lower meat quality. Several variants affecting the *SCG3* gene have been associated with obesity in human (Tanabe et al. 2007), supporting its likely causality for the fat-associated phenotypes in pigs.

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

A QTL on chromosome 1 affects growth and backfat in the Duroc breed (Table 8.1). The lead SNP in the GWAS result is located at position 1:265,017,724. The analysis returns 706 variants in high LD with the lead SNP (Table S9). The second pCADD-ranked SNP (1:g.263595807G>T), a single missense variant affecting the *COPS4* gene is identified as likely causal. The variant substitutes an alanine for an aspartic acid at position 252 in the COP9 signalosome complex subunit 4 protein (ENSSSCP00000056478:p.Ala252Asp). The Asp252 allele is associated with less backfat and slower growth. Variants affecting *COPS4* have been associated with increased body weight in mice (Blake et al. 2017).

8.2.4.4 Balancing selection for causal variants in the breeding program

Several identified variants exhibit pleiotropic effects for important selection traits, e.g. variants affecting *HMGA1*, *SCG3*, *COPS4*, and *MC4R* (Table 8.1). Variants that positively affect backfat often have negative consequences for growth, while variants that positively affect intramuscular fat often show detrimental effects on

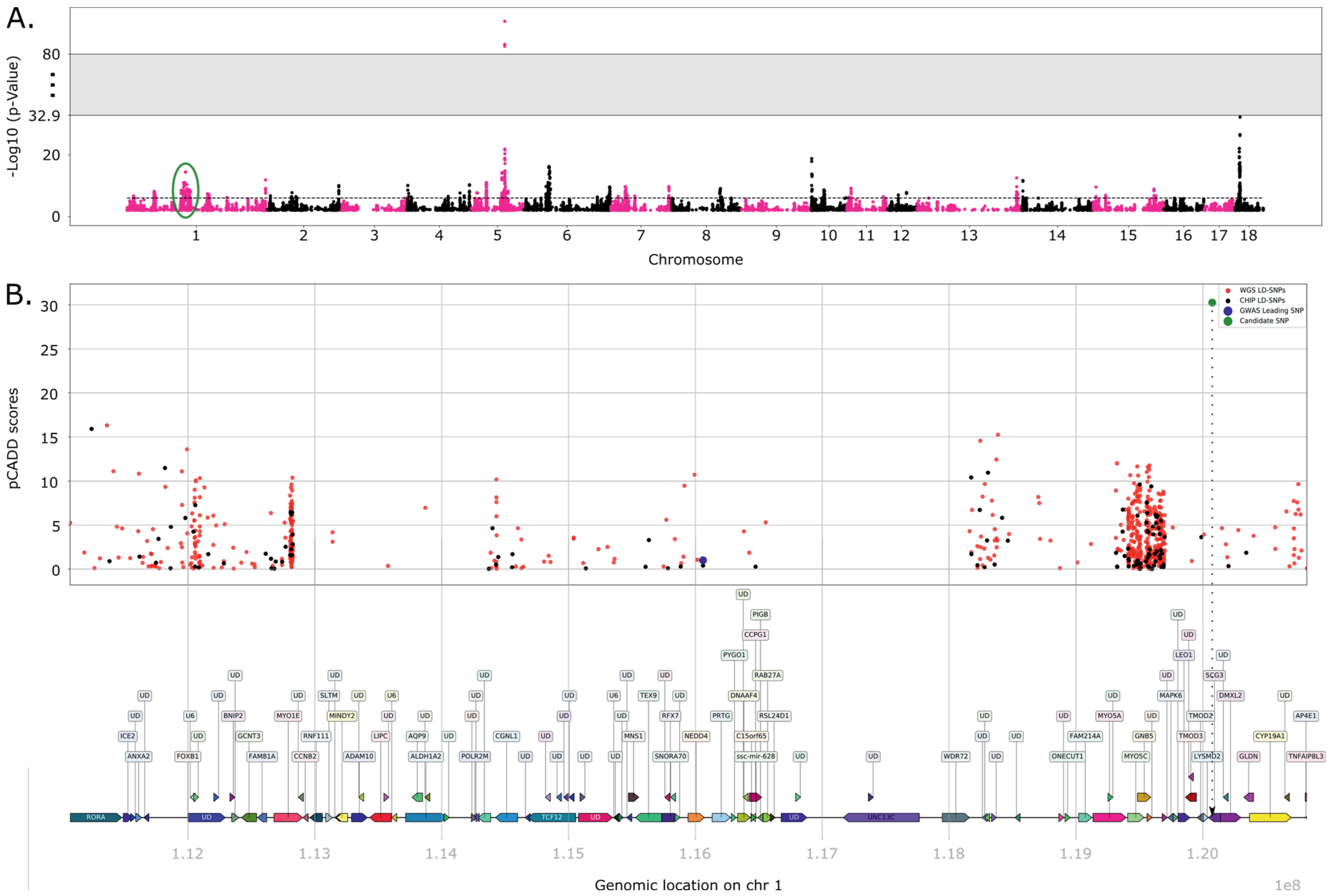


Figure 8.4 A) Manhattan plot for backfat in the Synthetic breed showing a strong QTL on chromosome 1:116Mb. Only SNPs with a -log$_{10}$(p) > 2 are plotted. B) Plot showing all sequence variants in high LD (red) with the lead SNP (blue), including the variants that are already on the chip (black), and the candidate causal variant (green). The bottom of the figure shows the gene annotation and location of the candidate causal variant, according to the Ensembl pig build v.98.

meat quality. The observed pleiotropic effects cause the variants to be under balancing selection in the breeding program, preventing population fixation of individual variants underlying strong QTL regions.

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

8.2.5.1 Genes affecting meat quality involved in muscle glycogen storage

We identified several candidate causal variants that affect meat quality. Especially in the Synthetic breed we find 26 loci significantly associated with drip loss (-log$_{10}$(p) > 6), a meat quality trait that measures the water holding capacity of the meat (Figure 8.5). The top ranked pCADD-scored genes show a strong enrichment for pathways involved in glycogen synthesis and storage (Table 8.1). Increased levels of muscle glycogen leads to increased drip loss, negatively affecting meat quality (Rosenvold et al. 2001). Examples of such variants include two regulatory variants affecting the *MEF2C* and *GBE1* genes. *MEF2C* knockout mice accumulate glycogen in their muscles (Anderson et al. 2015), while *GBE1* codes for a glycogen branching enzyme associated with glycogen storage disease, if mutated (Froese et al. 2015). Moreover, we identify two missense variants affecting the *NEU3* (ENSSSCP00000034065:p. Pro419Ser) and *MAP1A* (ENSSSCP00000005070:p.Gly1904Ser) genes, both directly involved in the glycogen deposition (Halpain and Dehmelt 2006; Yoshizumi et al. 2007).

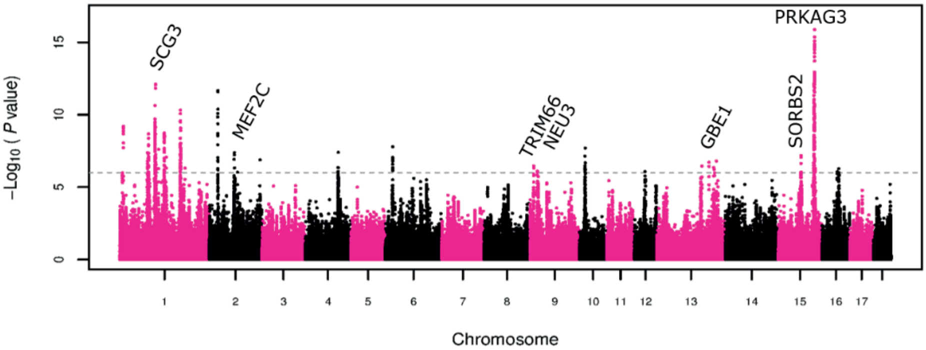


Figure 8.5 Manhattan plot for drip loss in the Synthetic breed. The figure shows significant loci and likely causal genes identified.

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

We identified several likely causal variants and genes affecting other important production traits (Table 8.1). The top-ranked genes are enriched in energy reserve metabolic processes, glycogen metabolic process, regulation of lipid biosynthetic process, and homeostasis (Table S10). More specifically, two identified regulatory variants in the *SOD1* and *PRKCE* genes likely affect backfat. *SOD1* is involved in glucose metabolism, and prevents oxidative damage associated with obesity (Liu et al. 2013), while mutations in *PRKCE* decrease the amount of body fat (Castrillo et al. 2001). Furthermore, we identified one regulatory variant in the *CACUL1* gene affecting intramuscular fat. This gene inhibits adipogenesis via the peroxisome proliferator-activated receptor γ (PPARγ) (Jang et al. 2017). In addition, two missense variants affect intramuscular fat via the *LNPEP* (ENSSSCP00000051249:p.Leu334Ser) and *ABCA12* (ENSSSCP00000058038:p.Gly1693Cys) genes. *LNPEP* attenuates diet-induced obesity in mice through increased energy expenditure, and decreases the amount of adipose tissue (Niwa et al. 2015), while the *ABCA12* gene plays an important role in lipid transport, affecting carcass fat content in pigs (Piorkowska et al. 2014). We further identified regulatory variants in the *NR1H3*, *NR1H4*, and *PRCP* genes, all likely affecting growth (Table 8.1). *NR1H3* and *NR1H4* are paralogous genes both involved in lipid homeostasis (Sinal et al. 2000; Zhang et al. 2016a), while reduced levels of *PRCP* expression promote obesity by regulating the α-melanocyte-stimulating hormone (α-MSH) that regulates feeding behaviour. Finally, we found a missense variant in the *SLC46A1* gene associated with increased intramuscular fat (ENSSSCP00000020843:Gly131Arg) in pigs, known to affect glucose and fat levels in knockout mice (Blake et al. 2017).

Table 8.1 List of potential causal variants identified from the pipeline. Table shows the variants type, potential overlap with promoter or enhancer region (from liver, (Villar et al. 2015)), the change in amino acid (for missense mutations) and the pCADD score for variants affecting one or more important selection traits (BFE: backfat, IMF: intramuscular fat, TGR: growth rate, DRY: drip loss, NTE: number of teats). The causal variant for genes in bold have already been reported in literature.

Chr	Variant	Type	Promotor/ Enhancer	Amino acid change	pCADD	Rank	Gene	Breed(s)	BFE	IMF	TGR	DRY	NTE	Supporting evidence
1	G-120074006-A	missense	NO	T386M	30.27	1	*SCG3*	Synthetic	-	+	NS	-	NS	Associated with obesity (Tanabe et al. 2007).
1	G-160773437-A	missense	NO	D298N	27.47	1	***MC4R***	Synthetic, Duroc	-	NS	+	NS	NS	Associated with fatness, growth, and feed intake traits (Kim et al. 2000).
7	G-30318881-A	upstream	YES	-	14.96	1	*HMGA1*	Duroc	-	-	+	+	NS	Associated with pig growth and fat deposition traits (Kim et al. 2006).
5	T-30187091-C	upstream	NO	-	19.44	2	*HMGA2*	Synthetic	-	NS	NS	NS	NS	*HMGA2* deficiency in pigs leads to dwarfism (Chung et al. 2018).
15	C-120865869-T	missense	NO	E47K	26.37	1	***PRKAG3***	Synthetic, Duroc	NS	NS	NS	-	NS	A combination of two variants in *PRKAG3* is needed for a positive effect on meat quality in pigs. (Uimari and Sironen 2014)
6	C-67433001-T	intron	YES	-	11.87	2	*KLHL21*	Landrace	NS	NS	NS	+	NS	Affects creatinine levels in mice (Blake et al. 2017).
1	C-127921686-T	missense	NO	G1904S	23.03	1	*MAP1A*	Synthetic	NS	NS	NS	-	NS	Involved in glycogen synthesis (Halpain and Dehmelt 2006).
2	C-96202720-T	intron	NO	-	17.86	2	*MEF2C*	Synthetic	NS	NS	NS	-	NS	*MEF2C* skeletal muscle knockout mice accumulate glycogen in their muscle (Anderson et al. 2015).
9	C-9329652-T	missense	NO	P419S	20.91	3	*NEU3*	Synthetic	NS	NS	NS	-	NS	Overexpression increases glycogen deposition (Yoshizumi et al. 2007).
13	G-173634576-A	upstream	YES	-	15.68	1	*GBE1*	Synthetic	NS	NS	NS	+	NS	Glycogen branching enzyme (Froese et al. 2015).
9	G-758928-A	missense	NO	A773T	21.92	1	*TRIM66*	Synthetic	NS	NS	NS	-	NS	*TRIM66* is involved in regulating glycogen synthesis (Fan et al. 2019).
14	T-107058908-C	intron	YES	-	24.50	1	*SORBS1*	Synthetic	NS	NS	NS	-	NS	Glycogen binding protein (Nagy et al. 2018).
15	A-46758359-G	intron	YES	-	11.73	4	*SORBS2*	Synthetic	NS	NS	NS	-	NS	-
7	A-97614602-C	upstream	NO	-	11.95	2	***VRTN***	Duroc, Landrace, Large White	NS	NS	NS	NS	+	Associated with increased number of vertebrae in pigs (van Son et al. 2019).
8	A-102781174-G	missense	NO	M165V	21.27	1	*QRFPR*	Synthetic	NS	NS	NS	NS	-	The G-protein-coupled receptor *QRFPR* regulates bone formation (Baribault et al. 2006).
18	T-10098558-C	intron	NO	-	16.41	1	*HIPK2*	Synthetic	-	-	NS	NS	NS	Essential regulator of white fat development (Sjolund et al. 2014).
2	A-144841051-C	intron	NO	-	10.65	2	*NR3C1*	Synthetic, Large White	+	+	NS	NS	NS	Glucocorticoid receptor (Reyer et al. 2013).
5	G-65814519-A	missense	NO	V850I	23.10	1	*AKAP3*	Duroc	+	NS	NS	NS	NS	Candidate gene for meat tenderness traits (Casiro et al. 2017).
11	T-20619202-C	3'UTR	NO	-	18.46	1	*HTR2A*	Duroc	-	NS	NS	NS	NS	Positive role of *HTR2A* in adipogenesis (Yun et al. 2018).
13	A-195332161-G	intron	YES	-	6.13	26	*SOD1*	Duroc	-	NS	NS	NS	NS	Associated with abnormal body fat mass (Liu et al. 2013).
3	C-94863278-A	5'UTR	YES	-	16.11	7	*PRKCE*	Landrace	-	NS	NS	NS	NS	Associated with decreased total body fat amount (Castrillo et al. 2001).
14	G-128748846-A	5'UTR	YES	-	15.53	7	*CACUL1*	Synthetic	NS	-	NS	NS	NS	Exhibits a repressive role in PPARγ activation and fat accumulation (Jang et al. 2017).
2	T-103610859-C	missense	NO	L335S	21.45	1	*LNPEP*	Synthetic	NS	+	NS	NS	NS	Decreased white fat cell size, decreased susceptibility to diet-induced obesity (Niwa et al. 2015).
2	C-41019232-T	upstream	NO	-	3.91	9	*SAA3*	Large White	NS	+	NS	NS	NS	Decreased susceptibility to diet-induced obesity, increased white fat cell size (den Hartigh et al. 2014).
4	A-88412353-C	intron	NO	-	18.99	1	*NOS1AP*	Large White	NS	+	NS	NS	NS	Regulates glucose homeostasis and hepatic insulin sensitivity in obese mice (Mu et al. 2019).
15	C-117292901-A	missense	NO	G1693C	24.75	1	*ABCA12*	Large White	NS	+	NS	NS	NS	Associated with pig production traits (Piorkowska et al. 2014).
6	A-146830209-G	intron	NO	-	11.88	1	*LEPR*	Duroc	NS	+	NS	NS	NS	The gene expression results showed that in the loin muscle LEPR showed significantly higher expression in pigs with higher IMF% (Li et al. 2010).
7	A-11391274-G	intron	NO	-	9.72	1	*JARID2*	Duroc	NS	+	NS	NS	NS	Regulates cell-cycle in skeletal muscle (Adhikari et al. 2019).
2	G-15310202-A	3' UTR	NO	-	21.38	1	*NR1H3*	Synthetic	NS	NS	-	NS	NS	Obesity, associated with lipid deposition in pigs (Zhang et al. 2016a).
5	A-83681067-G	intron	YES	-	12.10	9	*NR1H4*	Synthetic	NS	NS	-	NS	NS	Glucose tolerance, lipid homeostasis (Sinal et al. 2000).
9	C-17077403-A	5' UTR	YES	-	8.45	7	*PRCP*	Synthetic	NS	NS	-	NS	NS	Reduced levels of PRCP promote obesity (Palmiter 2009).
12	C-44684331-G	missense	NO	G131R	25.81	1	*SLC46A1*	Synthetic	NS	NS	+	NS	NS	Decreased total body fat amount, decreased circulating glucose levels (Blake et al. 2017).
6	A-145977262-T	intron	NO	-	14.63	1	*SGIP1*	Large White	NS	NS	-	NS	NS	Suppression of SGIP1 reduced body weight (Trevaskis et al. 2005).
1	G-263595807-T	missense	NO	A252D	20.05	2	*COPS4*	Duroc	+	NS	-	NS	NS	Increased body weight in KO mice (Blake et al. 2017).

8.3 Discussion

The aim of this study was to prioritize variants associated with important traits in pigs. The variants are ranked based on pCADD scores, and possibly further supported with respect to their function by epigenetic marks and gene expression data. The method is especially relevant because genomic variation underlying phenotypic variation mostly affects the non-coding part of the genome (Ponting and Hardison 2011), and GWAS results often point to regions outside gene boundaries (Bartonicek et al. 2017). With the publication of the pCADD scores (Gross et al. 2020), a powerful resource is now available to rank any possible substitution variant in the genome based on the likelihood of being functional. This is a major step forward in livestock, as thus far only variation in the coding region could be scored. On top of the pCADD scores, we use epigenomics and gene expression data to annotate regulatory sequences and associate gene expression to the trait of interest. In human, many transcriptomic and epigenomic marks have already been incorporated in the CADD scores (Rentzsch et al. 2019). However, the pCADD scores are built on far less (epi)-genomics data, but with the accumulation of functional genomic data in pigs (Giuffra et al. 2019), these pCADD scores will further improve.

Livestock populations generally have small effective population sizes (N_e: 50-200), far less compared to e.g. human (N_e ~ 10,000), leaving much longer blocks of variants in high LD. This high level of LD increases the power to detect QTL regions, even with relatively low SNP density. However, within large LD blocks, many variants will be associated, and a thorough variant prioritization should be performed to point to likely causal variants within the (often) large variant set. For example, the LD block for number of teats in Landrace spans about 1.8 Mb, leaving many thousands of variants in linkage, which increases the level of noise and hampers the detection of the causal variant. Nevertheless, in Large White and Duroc, which have smaller LD-blocks (100-500 kb), the causal *VRTN* promoter SNP is among the top SNPs. In that sense, integrating the results from multiple breeds provides additional power to further narrow down the list of candidates, assuming that the same causal variant is segregating, but likely with a very different underlying haplotype structure. This example shows that the tool can be very powerful to prioritize variants, but with a trade-off for the level of LD, increasing the noise if many thousands of variants are in linkage.

Although the development of genomic selection has revolutionized the world of animal breeding, the lack of functional genomic information currently limits further development (Georges et al. 2019). The framework and associated pCADD scores provided within this study will accelerate the discovery of new functional variants, which can be directly implemented in genomic selection by adding the causal variants to the selection chip used for genomics selection. Moreover, the results provide further knowledge of the biological pathways associated with important phenotypic variation in livestock. For this, the (functional) genome annotation in livestock genomes is still of too low quality compared to other well-studied mammalian species (Giuffra et al. 2019). Therefore, using annotations from human, mice, and rat will often provide more detailed information on gene function, pathways, and associated genes compared to the pig annotation itself.

The populations under study provide an interesting framework to study common pathways and molecular mechanisms involved in comparable phenotypes between pig and human. For example, we report the *GBE1* gene affecting meat quality in pigs by accumulating glycogen in the muscle, a gene associated with glycogen storage disease in human (Bao et al. 1996). Moreover, several of the identified genes affecting growth and fat deposition traits in pigs are involved in energy metabolism, glucose homeostasis, and adipogenesis, often associated with metabolic disease in human (e.g. *HMGA1*, *SCG3* genes). In human, however, environmental factors play a very large role in the formation of metabolic disease, while in pigs the animals are kept under relatively stable conditions, which could make the pig an ideal model to study the effects of specific genic variants on these analogous phenotypes (Perleberg et al. 2018). Pig breeding has led to extreme changes in animal production and efficiency, with very little negative consequences on health (Knol et al. 2016). This remarkable robustness of the animals, and the molecular mechanisms involved, could help to understand metabolic disease in human. Finally, our study implicates that, despite the complexity of pathways, there are several key entry points (i.e. genes) with a large effect on specific phenotypes in pigs, likely to be similar in human. Understanding these 'key' genes, and how they function together would further help to unravel the (molecular) consequences of genomic selection.

8.4 Conclusion

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

8.5 Methods

8.5.1 Ethics statement

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

8.5.2 Genotype data and breeds

The dataset consists of 15,791 (Duroc), 28,684 (Synthetic), 36,956 (Large White), and 41,865 (Landrace) animals genotyped on the (Illumina) Geneseek custom 50K SNP chip with 50,689 SNPs (50K) (Lincoln, NE, USA) and imputed to the Axiom porcine 660K array from Affymetrix (Affymetrix Inc., Santa Clara, CA, United States). The chromosomal positions were determined based on the Sscrofa11.1 reference assembly (Warr et al. 2019). SNPs located on autosomal chromosomes were kept for further analysis. Next, we performed per-breed SNPs filtering using following requirements: each marker had a MAF greater than 0.01, a call rate greater than 0.85, and an animal call rate > 0.7. SNPs with a p-value below $1x10^{-12}$ for the Hardy-Weinberg equilibrium exact test were also discarded. All pre-processing steps were performed using Plink v1.90b3 (Purcell et al. 2007).

8.5.3 Phenotypes

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

8.5.4 Genome wide association study

A single SNP GWAS was performed with the software ASReml (Gilmour et al. 2009) by applying the following model:

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

where DEBV ij is the DEBV (deregressed estimated breeding value) for genotyped animal j, µ is the overall DEBV mean of the genotyped animals, SNP_i is the genotype of the SNP *i* coded as 0, 1 or 2 copies of one of the alleles, a_j is the additive genetic effect and e_{ij} the residual error. The weighting factor w was used in the GWAS to account for differences in the amount of available information on offspring to estimate DEBV (Garrick et al. 2009). Association results were considered significant if -log10(p) > 6.0.

8.5.5 Population sequencing and mapping

Sequence data was available for 101 (Duroc), 71 (Synthetic), 167 (Landrace), and 89 (Large White) animals from paired-end 150 bp reads sequenced on Illumina HiSeq. The sequenced samples are frequently used boars, selected to capture as much as possible of the genetic variation present in the breeds. The coverage ranges from 6.6 to 22.2, with an average coverage of 11.82 (Table S12). FastQC was used to evaluate read quality (Bioinformatics 2011). BWA-MEM (version 0.7.15, (Li and Durbin 2009)) was used to map the WGS data to the Sscrofa11.1 reference genome. Samblaster was used to discard PCR duplicates (Faust and Hall 2014), and samtools was used to merge, sort, and index BAM alignment files (Li et al. 2009).

8.5.6 Variant discovery functional class annotation

Freebayes was used to call variants with following settings: --min-base-quality 10 --min-alternate-fraction 0.2 --haplotype-length 0 --ploidy 2 --min-alternate-count 2 (Garrison and Marth 2012). Post processing was performed using bcftools (Li et al. 2009). Variants with low phred quality score (<20), low call rate (<0.7) and variants within 3 bp of an indel are discarded, leaving a total of 21,648,132 (Landrace), 23,667,234 (Duroc), 23,286,212 (Synthetic), and 25,709,552 (Large White) post-filtering variants, respectively. The average per variant call rate is above 98% for all breeds and the ratio transitions to transversions is between 2.33-2.35 (Table S11). Variant (SNPs, Indels) annotation was performed using the Variant Effect Predictor (VEP, release 97) (McLaren et al. 2016).

8.5.7 pCADD scores

pCADD scores were retrieved from Gross et al (2019). Visualization of pCADD scores was performed using JBrowse 1.16.6 (Skinner et al. 2009). Integration of sequence variants with pCADD score was performed using PyVCF (Casbon 2012). pCADD scores, partitioned per chromosome, compressed via bgzip and tabix indexed for fast access, can be downloaded following this link (~5GB-1GB): http://www.bioinformatics.nl/pCADD/indexed_pPHRED-scores/, and scripts to use these scores to annotate SNPs can be found here: https://git.wur.nl/gross016/pcadd-scripts-data/.

8.5.8 Promoter and enhancer elements from ChipSeq data.

We retrieved three H3K27Ac, and three H3K4me3 libraries (ArrayExpress accession number: E-MTAB-2633) from liver tissue from three male pig samples described by Villar et al. 2015. Data was aligned using BWA-mem (Li and Durbin 2009) and visualized in JBrowse (Skinner et al. 2009). Coverage information on variant sites was obtained using PyVCF (Casbon 2012) and the PySAM 0.15.0 package.

8.5.9 Phenotypes and gene ontology.

Phenotype information from genes orthologous to pig in human, mouse, and rat were retrieved from the Ensembl database ((Hunt et al. 2018), release 97) using a custom bash script. Gene ontology and pathway information was obtained from the UniProt database (UniProt 2019).

8.5.10 RNA-sequencing and differential expression

We used 25 Duroc and 34 Landrace RNA-sequenced boars selected based on high and low sperm DNA fragmentation index, a measure of well packed double-stranded DNA vs single-stranded denatured DNA, which is an important indicator of boar fertility (van Son et al. 2017b). The boars were all born in the same period of time and a broad range of semen quality tests were conducted on ejaculates of these boars. Sequencing was done in two batches. Library preparation and sequencing strategy of the first batch can be found in van Son et al. 2017. The second batch was prepared using TruSeq mRNA stranded HT kit (Illumina) on a Sciclone NGSx liquid automation system (Perkin Elmer). A final library quality check was performed on a Fragment Analyser (Advanced Analytical Technologies, Inc) and by qPCR (Kapa Biosciences). Libraries were sequenced on an Illumina HiSeq 4000 according to manufacturer’s instructions. Image analysis and base calling were performed using Illumina’s RTA software v2.7.7. The resulting 100 basepair single-end reads were filtered for low base call quality using Illumina’s default chastity criteria. We mapped the RNA-seq data to the Sscrofa11.1 reference genome using STAR (Dobin et al. 2013) and called transcripts and normalized FPKM expression levels using Cufflinks and Cuffnorm (Trapnell et al. 2013). We assigned the genotype class (homozygous reference, heterozygous, homozygous alternative) for each RNA-sequenced individual using the 660K genotype of the lead SNP in the GWAS result. We tested for differential expression between three genotype classes using the one-way

ANOVA test. The Welch t-test was used to evaluate the differences between two genotype classes. A p value < 0.05 was considered significant.

8.6 Additional files

All supplementary material are available at the Open Science Framework repository: https://osf.io/cyu2m/

8.7 Acknowledgements

The authors thank Egbert Knol, Roel Veerkamp, Marco Bink, and Barbara Harlizius for useful input on this work.

8.8 Authors' contributions

MG, H-JM, DdR, MR, and MD conceived and designed the study. MD and CG performed the data analysis and wrote the manuscript. ML performed GWAS analysis. H-JM, MG, MB, MR, DdR, AG, ML provided useful comments and suggestions and helped to draft the manuscript. All authors read and approved the final manuscript.

8.9 Competing interests

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

9

General discussion

9.1 Introduction

The genome of modern domesticated animals is shaped by a long history of selection (Wang et al. 2014). Despite the long selection history, we are only starting to understand the relationship between the variation in the genome and the phenotype in animals. In this thesis I explore the genomic variation currently present in commercial livestock breeds, and try to further bridge the genotype-phenotype gap by using novel methodologies.

In animal breeding, less priority has been given to functional and molecular genomics compared to a traditional statistical quantitative genetic approach. The realised genetic gains in commercial livestock populations can be mainly attributed to genomic selection, with very little knowledge on which genomic variation affects the selection traits. However, the use of genomic selection coincided with a vast increase of ~omics data for breeding purposes. Hundreds of thousands of animals have now been genotyped (mostly on medium-density SNP arrays) in elite breeding populations, complemented with hundreds of whole genome sequenced animals. This large volume of data provides new opportunities to identify important alleles in the population, either deleterious or beneficial, using a set of different methodologies. In this thesis I present tools to identify such high-impact alleles in the breeding populations, and I discuss how to efficiently use the information for breeding purposes.

In the first part of my thesis, I emphasized on harmful genetic variation (i.e. lethal recessives) present in the populations as a consequence of selection and inbreeding. I described in detail the mutation, molecular consequence, and associated phenotypes of several lethal variants in various livestock breeds, which contributes to our general understanding of genetic defects in livestock. In addition, I studied in detail different types of deleterious variants in commercial breeding populations, and discussed the effect of genetic drift, purifying selection, and recombination, shaping the (local) deleterious landscape of the genome. In the final part of my thesis, I go beyond deleterious, emphasizing on the identification of functional variation, underlying important selection traits. We provide a toolbox that can enhance the identification of causal (regulatory) variation in livestock populations. In this chapter, I discuss the implications and main conclusions from the chapters, and how this thesis contributes to the current literature. Finally, I discuss the applications, opportunities, and future perspectives this thesis provides for animal breeding.

9.2 Deleterious alleles

Deleterious alleles cause a decrease in fitness compared to the effects of other usually more common alleles in the population. The effects of the deleterious alleles depends on the evolutionary and selection history of the population (Charlesworth and Willis 2009). Over time, deleterious alleles are purged from a population by (natural) selection. This purging is very efficient for dominant deleterious alleles that lower the fitness of carrier animals. However, as a function of the low frequency, recessive deleterious alleles are generally masked from natural selection by a dominant non-deleterious allele, and will therefore be easily passed on into the next generation.

Most studies aiming to identify deleterious alleles from sequence data are focussing on missense and loss-of-function variants in the coding regions of the genome. However, accurate prediction of deleterious alleles from sequence data is not straight forward. This is especially true for regions with lower sequence mapping quality, or for homologous regions in the genome (e.g. for large gene families), that are prone to produce many false positive variant calls driven by spurious alignments (Nielsen et al. 2011). Stringent filtering criteria can be applied to reduce the number of false positives, which can also be a result of errors in the functional annotation. However, even after stringent filtering, the final list of deleterious and loss-of-function variants should be handled with caution, especially for making functional predictions on individual variants. In addition, the majority of the deleterious variants are likely of regulatory origin (Rojano et al. 2019), not affecting the coding sequence directly. Hence, making accurate functional predictions for variants outside the coding region is even far more challenging.

In chapters 4 and 6 I investigated in detail the landscape of deleterious alleles in commercial chicken and pig populations. The identification of deleterious alleles still relied on “traditional” methods that annotate the genomic variation using the Ensembl Variant Effect Predictor (McLaren et al. 2016), and predicting the deleteriousness of missense variants with SIFT software (Kumar et al. 2009). In this thesis, I show clear evidence of purifying selection acting on deleterious variation (including loss-of-function variation) inferred from an altered frequency distribution for deleterious variants. In addition, purging is more effective in regions with higher recombination rate (chapter 4 and 6). Higher rates of recombination likely cause deleterious variants to be more easily separated from positive variation, enhancing purifying efficiency, and leading to a more diverse landscape of (recombinant) haplotypes (Bosse et al. 2018). The effect is evident both in pigs and chicken, which have very different recombination landscapes across their genomes (Megens et al. 2009; Tortereau et al. 2012).

The complete set of deleterious alleles within a genome can be used to assess the “genomic fitness” and infer the level of inbreeding within a population. The effects of deleterious alleles vary. Some have large effects (most notably recessive lethals), but the majority of the “deleterious load” is caused by the cumulative outcome of many deleterious alleles with small effects (Charlesworth and Willis 2009). The genomic fitness can be determined by the ratio of deleterious variation over neutral variation in a population (chapter 4; (Bosse et al. 2018)). Another well-established method to quantify inbreeding is to assess the overall homozygosity of the genome. This can be inferred by quantifying homozygous stretches in the genome called runs of homozygosity (ROH). ROHs are enriched for homozygous deleterious variants (Bortoluzzi et al. 2019), and therefore is useful as indicator for inbreeding risk. While both methods can provide accurate estimates of the deleterious load, it usually does not provide any lead to the effects of any particular deleterious variants, nor does it indicate which variants contribute to the negative fitness effects in inbred populations (if present). For example, only several hundreds of variants show a deleterious effect in livestock, reported in the Online Mendelian Inheritance in Animals (OMIA) database (Nicholas 2003). Many more exist, but the lack of impact prediction tools, especially for the non-coding regions of the genome, has hampered the identification of deleterious variation. However, the published Combined Annotation Dependent Depletion (CADD) scores in livestock now provide new opportunities to disentangle variation with impact (i.e. contributing to the negative fitness effect) and variation that is benign (Gross et al. 2020). The CADD tool, originally developed in human (Rentzsch et al. 2019), is built on many layers of annotations, including conservation scores, sequence context, and (epi)-genomic features. Subsequently, the machine learning method generates a set of proxy-benign variants that became fixed or nearly fixed in the pig lineage compared to an ancestral sequence, surviving millions of years of purifying selection (likely enriched for neutral variants). Next, a second set of simulated proxy-deleterious variants is generated, of which a considerable fraction would likely be deleterious. The contrast between both variant sets (proxy deleterious vs. proxy benign) and the differences in their features provides the core characteristic of CADD. The CADD scores now allow for more accurate deleterious load and population fitness predictions compared to traditional methods, that mostly rely on the coding sequence.

9.3 Lethal recessives

In this thesis I have delved into recessive lethal variation in livestock. I describe different types of mutations, affecting genes essential for normal development, resulting in embryonic, fetal, or postnatal mortality in homozygous animals (Figure 9.1) (Derks et al. 2018; Derks et al. 2019). The identification relies on pinpointing haplotypes in the genome are never found in homozygous state, or that occur far less often in homozygous state than expected. The method relies on two assumptions, 1) the lethal recessive variant should be in high LD with the haplotype, and 2) the frequency and quantity of genotypes should be sufficient to prove a statistically significant lower occurrence of homozygotes in the population. Though this method has proven very successful in identifying lethal recessive variants, many lethal recessives residing on more common haplotypes (i.e. variants that emerged more recently) cannot be identified in this manner. For example, Charlier et al showed that, using this method, only about 25% of the embryonic lethal variants could be identified in New Zealand dairy cattle (Charlier et al. 2016), suggesting that a large reservoir of recessive lethals remains undetected.

After haplotype identification, the phenotypic effect can be assessed in carrier-by-carrier crosses (carriers exhibit a single copy of the lethal variant). An interesting observation in chapter 6 is that none of the lethal recessives described led to the (theoretically) expected litter size decline of 25% in carrier-by-carrier matings (the decline was mostly between 15-20%). I believe the lower-than-expected mortality is likely caused by a surplus of zygotes in the uterus of the sow, exceeding the uterus capacity. This phenomenon is likely even more relevant for homozygotes that lead to early (embryonic) lethality, leaving room for viable blastocysts to implant and develop, though this hypothesis will remain difficult to prove.

The current enormous amount/volume of genotype information generated through routine breeding practices in livestock allows for additional methods to trace deleterious haplotypes in the genome. One avenue for research is to assess the inheritance patterns of specific haplotypes over time, which could provide evidence for purifying selection. Another promising method to identify deleterious alleles in the population is to investigate transmission ratio distortion of alleles, i.e. the deviation from the expected mendelian ratio in offspring (Casellas et al. 2017).

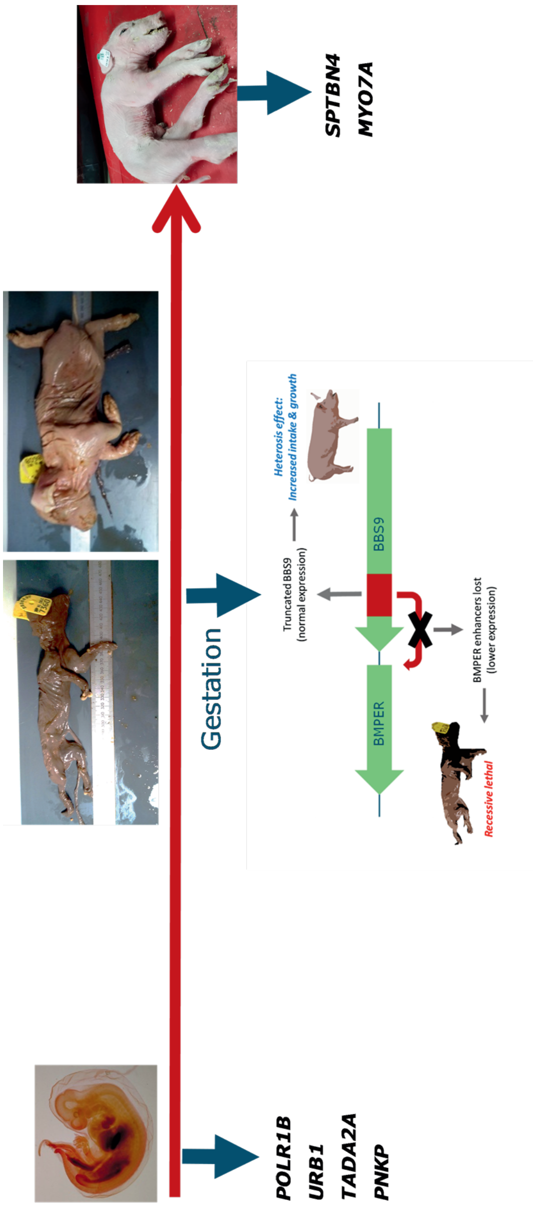

POLR1B
URB1
TADA2A
PNKP
Gestation
Truncated BBS9 (normal expression)
Heterosis effect: Increased intake & growth
BBS9
BMPER
BMPER enhancers lost (lower expression)
Recessive lethal
SPTBN4
MYO7A

Figure 9.1 Overview of recessive lethals described in this thesis. Shown are the effected genes, and the developmental state , from early embryo until post-partum, at which homozygous animals die.

9.3.1 Finding the needle in the haystack using population ~omics data

Finding the causal variant, even within a relatively small haplotype region, is not trivial. First of all the quantity of sequenced animals in the population should be sufficient to identify carrier animals, that then are instrumental for determining linkage disequilibrium (LD), i.e. the correlation between the haplotype and the deleterious sequence variants. Loss of function and deleterious missense variants in high LD with the haplotype are often of first interest, because these are likely to disrupt gene function. However, WGS data alone is not always sufficient to identify the causal variant, even in these seemingly evident cases. In chapter 6, for example, I report two variants that affect intron-exon splicing, either by skipping a complete exon or by the retention of a complete intron. Either consequences usually result in the production of an erroneous mRNA that contains premature termination codons, leading to a loss-of-function of the protein. Although the splicing dinucleotides (GT-AG) are extremely conserved, variants in these sites do not always lead to miss-splicing (Lewandowska 2013). Hence, including RNA-sequencing data is crucial to verify the consequence of the splice mutations, which is even more relevant for splicing variants that do not affect the splice dinucleotides, but other sites within the splicing region (e.g. the *POLR1B* splice variant described in chapter 6).

Alternatively, the causal variant can be structural in nature, i.e. be a deletion or inversion of part of a gene or regulatory region. In chapter 5, I describe an unique example of allelic pleiotropy, with a large deletion affecting two neighbouring genes. While the detrimental effect of the deletion on the *BBS9* gene seemed immediately evident, this was not the case for the effect on the downstream *BMPER* gene. Hence, the consequence of the deletion could only be assessed by including publicly available ChIP-seq data (H3K27Ac, H3K4Me3), marking active regulatory promoter and enhancer regions in the porcine liver (Villar et al. 2015). Further evidence supporting the miss-regulation of the *BMPER* gene, as a consequence of the deletion, was inferred from allele specific expression of the wild-type haplotype in a carrier animal, i.e. the *BMPER* gene has lower expression on the deletion haplotype.

The increasing volume of population whole genome sequence data opens up possibilities to predict deleterious variants directly from sequence, without prior knowledge on the haplotypes inferred from genotype data. Large scale population WGS data also helps to separate noise from “true” deleterious variants. Hence, the list of deleterious variants inferred from the population sequence data will thereby become more reliable, allowing for the identification of high-impact variation purely from sequence. For example, the deleterious effect of the *SPTBN4* frameshift deletion was purely derived from sequence, without prior haplotype knowledge. The identified deleterious variants can subsequently be monitored in the population, in order to prevent a rapid increase in the frequency. Ideally, the variants are placed on the selection chip (used for genomic selection) to validate its consequence in homozygous state, and to allow for efficient selection against it.

The phenotype of homozygous animals can provide further leads to the causal gene, especially if homozygotes die postnatally. For example, the phenotype of the *SPTBN4* knockout pigs (myopathy and tremors) is analogous to (natural) knock-out phenotypes in human and mice, supporting that the syndrome is caused by an impaired *SPTBN4* gene (chapter 7). I show that the genetic defects does express itself to breeders, but was never recorded in breeding records, likely caused by the lack of prior knowledge on farms regarding genetic defects in purebred populations. However most of the identified recessive defects lead to early (embryonic) lethality, described in chapter 6, leaving no trace except for the absence of homozygotes in the population.

9.3.2 Balancing selection or genetic drift

In this thesis I describe three different population genetic forces that can drive the frequency of deleterious variants: 1) genetic drift, 2) genetic hitchhiking with a beneficial variant, or 3) a direct beneficial effect of the deleterious variant in heterozygotes. The latter two can cause balancing selection of the variant in the population, driven by the positive effect in heterozygotes. Evidence for heterozygous advantage based on an association test between a group of carrier and non-carrier animals, because the frequency is likely too low to be picked up in a regular GWAS study. The reliability of the association test depends on the group sizes and the accuracies of the breeding values, which are mostly deregressed (indirect estimation of the phenotype) to estimate the actual phenotypic effect (Garrick et al. 2009). In this thesis, we show one clear example of balancing selection for the *BBS9* deletion with pleotropic effects on two different traits, resulting in a higher selection index for carrier animals (chapter 5). Although the result on the deletion might suggests a larger set of deleterious variants under balancing selection, we did not find any evidence for that. I even believe that most lethal variants are purely the result of genetic drift, at least that appears the most likely cause for relative high frequencies of some of the other identified recessive lethals in this thesis. Nevertheless, a possible positive association signal can be lost if the beneficial variant underlying the hitchhiking effect reaches population fixation (Charlesworth 2007).

9.3.3 Two cases of recessive defects in Duroc should be further investigated

Studying lethal recessives is an ongoing topic of research. For example, an additional screen for lethal recessives in a Duroc breed revealed two haplotypes that likely carry a lethal recessive variant on chromosome 9 (Table 9.1).

Table 9.1: Two haplotypes exhibiting a deficit in homozygosity in Duroc. Shown are the location, frequency, and observed vs. expected number of homozygotes.

Abr.	Chr	Pos (Mb)	Hap Frq.	Carriers	Obs. Homozygotes	Exp. Homozygotes
D1	9	121.5-122.5	0.106	3016	2	161
D2	9	11.5-12.5	0.042	1158	2	24

Case 1: A recessive defect segregates at high frequency in a Duroc population

The first haplotype (D1) has a carrier frequency of about 21%. For recessive lethals, I simulated that with N_e=100, the maximum carrier frequency is about 20% before the lack of homozygotes will prevent further increase (chapter 6). Hence, this high frequency is likely not purely caused by drift, but rather the result of a beneficial effect in heterozygotes.

The 308 identified carrier-by-carrier crosses show a 12.5% reduction in total number born (Table 9.2). This 12.5% reduction is highly significant, but falls outside the expected range (15-23%) observed for the other lethal recessive haplotypes (chapter 6). This rather low reduction might be caused by incomplete LD between the haplotype and the causal variant. However, this seems rather unlikely since only two homozygous animals were found, while 161 are expected. Hence, this difference suggests high LD between the haplotype and causal variant. Another hypothesis is that the causal variant leads to lethality before implantation, in which the compensatory effect of viable embryos might increase, resulting in fewer piglet losses.

Table 9.2: Fertility phenotypes for haplotype D1. Table shows the total number born (TBN), number born alive (NBA), and postnatal survival in carrier-by-carrier crosses (CxC), carrier by non-carrier crosses (CxN), and non-carrier by non-carrier crosses (NxN).

Status	Number	TNB	NBA	Farrowing survival (%)	Preweaning survival (%)
CxC	308	9.06*	8.36*	92.19	91.41
CxN	2005	10.36	9.47	91.05	89.80
NxN	3052	10.38	9.38	89.84	88.99

*P < 0.05

Whole genome sequence analysis of five carrier animals revealed one loss-of-function variant in complete LD with the D1 haplotype. The mutation is a 4bp frameshift insertion in the *SOAT1* gene (CG-121056051-CTTTTT). *SOAT1* (Acyl-coenzyme A:cholesterol acyltransferase) is an intracellular protein located in the endoplasmic reticulum that forms cholesterol esters from cholesterol, and has been associated with hyperlipidemia (Lu et al. 2011). However, homozygous knock-out mice (*SOAT1*) are viable, with abnormal blood lipid levels and increased circulating cholesterol levels (Yagyu et al. 2000). We therefore assume that the *SOAT1* frameshift insertion is not causal, suggesting an alternative variant to be causal.

Case 2: A natural knockout of the *MYO7A* gene in pigs provides a model for the Usher syndrome in humans

The D2 haplotype segregates with a 8.4% carrier frequency in the population and reduces preweaning survival in carrier-by-carrier crosses by 20% (Table 9.3).

Table 9.3: Fertility phenotypes for haplotype D2. Shown are the total number born (TBN), number born alive (NBA), and postnatal survival in carrier-by-carrier crosses (CxC), carrier by non-carrier crosses (CxN), and non-carrier by non-carrier crosses (NxN).

Status	Number	TNB	NBA	Farrowing survival (%)	Preweaning survival (%)
CxC	31	10.19	9.06	87.90	69.99*
CxN	777	10.31	9.28	89.28	90.41
NxN	4556	10.29	9.36	90.61	89.55

*P < 0.05

Similar to the frameshift deletion in the *SPTBN4* gene (chapter 7), this haplotype leads to piglet mortality within the first few weeks of age. Whole genome sequence analysis revealed a strong candidate stop-gained mutation (C-11280403-T) in complete LD with the D2 haplotype. The mutation affects the *MYO7A* gene, leading to a stopcodon at position 181-Q/*, resulting in an impaired and truncated MYO7A protein.

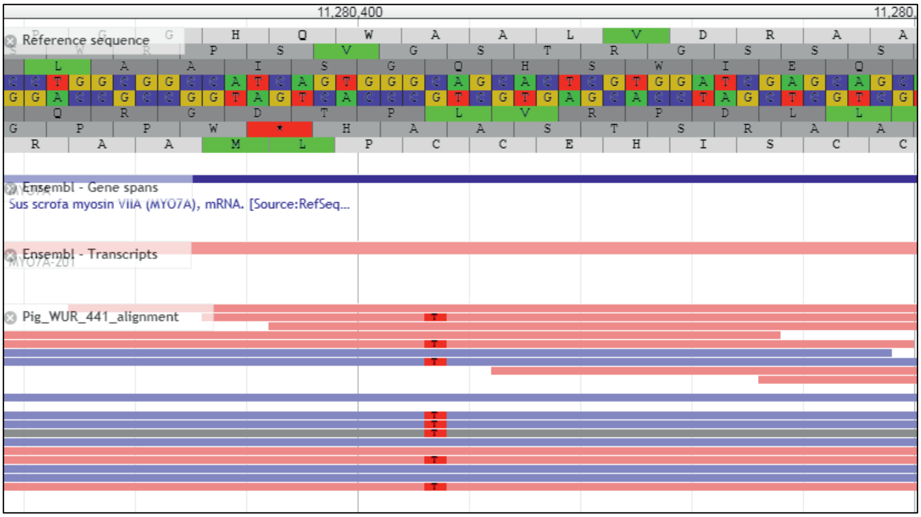


Figure 9.2: JBrowse screen capture of the *MYO7A*: C-11280403-T stop-gained mutation in one carrier animal. Figure shows the alignment of a D2 carrier animal within the coding region of the *MYO7A* gene.

Mutations in *MYO7A* are associated with the Usher Syndrome (deaf-blindness) in humans (Lentz and Keats 1993) and leads to head-shaking, deafness, retinal defects, and reduced male fertility in knock-out mice (Blake et al. 2017). Hence, this natural knockout could be used as a model to study the effects of an impaired *MYO7A* gene in pigs, and compare this to the human disease phenotype.

9.3.5 The contribution of lethal recessives to the heterosis effect

The functional genomic basis of the heterosis effect remains unresolved in breeding, i.e. fitness or other phenotypes score higher in offspring compared to the parents if derived from different populations or breeding lines (Charlesworth and Willis 2009). The heterosis effect is dominated by two main theories; 1) The dominance hypothesis emphasizes the suppression of undesirable recessive alleles (by dominant alleles); 2) The overdominance hypothesis emphasizes on heterozygote advantage, attributing heterosis to the superior fitness of heterozygous genotypes. This thesis shows that the suppression of recessive lethals in crossbred animals contributes to the heterosis effect for fertility (chapter 6), supporting the dominance hypothesis. Although significant, this effect contribution of lethal recessives is only marginal, and other detrimental, but not lethal variants, in purebred populations likely contribute to the heterosis effect even more.

9.3.6 Application in animal breeding

The majority of the recessive lethal variants are breed specific, not affecting the crossbred production animals, because no homozygotes will appear. Despite the small effect (if not zero) on production animals, carrier-by-carrier crosses in the purebred populations should be avoided, especially if homozygous animals die later in gestation (chapter 5) or postnatally (chapter 7). I can think of two main reasons to avoid such matings: 1) preventing the birth of affected individuals, improving animal welfare, 2) to avoid production losses in the nucleus farms. Topigs Norsvin implemented a mating-strategy pipeline to avoid such carrier-by-carrier crosses by identifying carriers in the selection animals, a strategy that has earlier been applied in several cattle breeds (Upperman et al. 2019).

Some of the current selection arrays are now designed with a flexible part, in which SNPs can be replaced annually (i.e. SNP arrays can be augmented with a set of custom SNPs designed by the breeding company). This flexible part allows for fast implementation of newly discovered deleterious variants on the selection array. The information derived from the genotype data for these putative deleterious variants can lead to validation of the functional consequence of the variant, and the fast identification of carrier animals can directly facilitate purging. Finally, the work described will raise awareness among farmers and breeders regarding “hidden” genetic defects in the populations. For example, in chapter 7, I describe a clear syndrome that was never specifically described in the breeding records, unless we notified the farmer to specifically monitor carrier-by-carrier crosses. The results will hopefully encourage farmers (especially on nucleus farms) to report and genotype complete litters that produce an interesting pattern of siblings that share a similar phenotype, seemingly affected by a genetic defect.

9.4 Accelerated discovery of functional variation in livestock

In this thesis I attempt to understand the relationship between the genotype and the phenotype of an individual. With the fast accumulation of population genotype, whole genome sequence, and expression data, we can more accurately assign function to variation (Yang et al. 2017, this thesis). More specifically, breeding companies now have the resources to identify almost all the variation present in a population. This data can be complemented with functional genomic information (e.g. on regulatory sequences), comparative data, and the CADD scores. However, the (functional) gene annotation in livestock genomes is still inferior compared to other well-studied vertebrate species (Giuffra et al. 2019). Therefore using comparative strategies, i.e. comparing annotations from well-studied species (e.g. human, mice, and rat) in an evolutionary framework, will often provide more detailed information on gene function, pathways, and associated phenotypes compared to the livestock annotation. With this complete set of population ~omics data, comparative strategies, and functional annotations, we can finally start identifying the functional consequences of variants with subtle impact, including variants that exhibit positive effects on traits important for breeding that usually are regarded as 'quantitative genetic' in nature.

9.4.1 Variant prioritization

Proper variant prioritization is key to assigning function to variation, aided by functional information of the genome. With the publication of the CADD tools, now available for pig and chicken (Gross et al. 2020), a powerful resource is available to rank any possible variant in the genome based on its likely importance, including the non-coding parts of the genome. In chapter 8 I clearly demonstrate the benefit of having a pCADD tool for variant prioritization, enhancing the identification of causal variants in genomic regions where associations are found with certain phenotypes. Recently, a chicken CADD has been developed, providing opportunities to subsequently perform similar analysis in chicken.

In livestock, these CADD scores are mainly built on sequence features, with sequence conservation as one of the most important features contributing the scores. The prediction accuracy of the CADD scores will further improve with the incorporation of functional (epi)genomic data, as was shown for the human CADD scores (Rentzsch et al. 2019). I predict that the Functional Annotation of ANimal Genomes (FAANG) consortium will play an essential role to annotate these functional genomic elements in livestock genomes (Giuffra et al. 2019). First, RNA-sequencing in various tissues will aid to identify transcribed loci in the genome, and histone modification marks, ATAC-seq, and DNA-methylation are used to assess the chromatin accessibility and architecture of the genome. In addition, Hi-C methods can be used to assess the 3D conformation of the genome. Hence, the public data, and methods generated within the FAANG consortium will be key to improve the CADD scores on the longer term.

9.4.2 Linkage disequilibrium, a friend and enemy.

Small effective population size and strong selection have led to a high degree of linkage disequilibrium in most commercial livestock populations (Hall 2016). This high degree of linkage between markers was useful in QTL analysis, as it allowed finding QTL even with low numbers and low SNP-density. However, high LD impedes the fine-mapping of the QTL region, often megabases in size. Fine mapping entails zooming in closer to the causal variant, which could be a single SNP, which means only one bp in size. High LD results in potentially thousands of variants to be similarly associated with a phenotype because of LD. One strategy is integrating the results from multiple breeds. This multi-breed approach provides additional power to further narrow down the list of candidates, assuming that the same causal variant is segregating, but likely with a very different underlying haplotype structure. Another possibility to overcome the LD problem is to use crossbred animals (Iversen et al. 2017), which are increasingly being genotyped and phenotyped. Crossbreeding causes fast breakdown of LD, allowing to more accurately finemap the QTL region and identify the causal mutation.

9.4.3 The identification of expression QTLs will enhance identification of functional variation

The most important factor that influences the phenotype of an animal is how the genes in the genome are regulated. Recent studies have shown that changes in gene expression, caused by changes in regulatory sequences in the genome, have a much larger impact on selection traits, whereas changes in the coding sequence only contribute sporadically (Ponting and Hardison 2011). Hence, variation in gene expression is the main factor influencing phenotypic variation in livestock. However, little attention has been given to gene expression in breeding practises, while this may be the major driver to understand how the regulation of the genome, and the molecular mechanism involved, affect the performance of an animal.

Expression QTL studies have become common practise in human genomics research (Schaid et al. 2018), but in livestock the amount of expression data is limited (Georges et al. 2019). While important selection sires are now regularly sequenced in livestock, to build a reference population that captures genomic variation present in the population, little is invested in RNA-sequencing. One limiting factor can be that the expression very much depends on the tissue, developmental stage, and the environment. Hence, the integration of expression data is not trivial, especially if the data comes from different stages, subject to various sampling/library preparation methods. In addition, currently more single-cell RNA-sequencing is applied, providing less biased results compared to pooled-based technologies. Nevertheless, future studies should focus more on the investigation of gene expression data in combination with whole genome sequence, genotype, and phenotype data to identify (non-coding) functional regulatory variation affecting genes underlying important traits. I propose to build a reference population of animals for which RNA is sequenced in various important tissues (trait dependent). This resource can be used to correlate genomic variation with changes in gene expression, and identify causal genes underlying important selection traits. The identified expression quantitative trait loci will greatly enhance our understanding how regulatory variation affects phenotypic variation in livestock.

9.4.4 Application in animal breeding

Although the development of genomic selection has revolutionized the world of animal breeding (Meuwissen et al. 2001), the lack of functional genomic information currently limits further development. Genomic selection works with a set of neutral SNP markers that are evenly distributed across the genome and particularly works well in closed elite-breeding populations. However, it is much less successful in diverse, multi-breed populations. A first attempt to increase marker density up to whole-genome sequence levels was not successful to improve breeding value prediction accuracies (van Binsbergen et al. 2015), mainly because millions of parameters (the SNP effects) must be estimated from a reference populations of only several thousand animals (“large p small n problem”). One proposed solution to improve prediction accuracies is to include the causal variants affecting the traits in genomic prediction (Perez-Enciso et al. 2015). This method of using weighted and pre-selected markers (i.e. functional markers) has proven very successful to improve prediction accuracies (Raymond et al. 2018b). Nevertheless, I predict that the within breed genomic prediction accuracies will likely not benefit that much from this functional information, since most of the causal variants will be captured by SNPs on the chip that are in high LD. However, for genomic prediction across breeds, having the causal variants will likely increase the genomic prediction accuracies substantially (Raymond et al. 2018a), due to strong differences in allele frequencies between populations.

This thesis provides tools that can accelerate the discovery of new functional variants (chapter 8). One major step to enhance the utilization of functional genomic information in animal breeding will likely be achieved by adding functional SNPs to the selection array (i.e. SNPchip used for genotyping and genomic selection). The fraction of functional SNPs on most current selection arrays is minute (Habier et al. 2013). However, with the adoption of functional variation in breeding, the number of SNPs predicting function will gradually increase.

Breeding companies now have started using SNP arrays of which part of the SNPs can be replaced annually, allowing for fast implementation of newly discovered variants in breeding. Using functional information for prioritizing variation increases genomic prediction accuracies substantially, as was also proven in cattle recently (Xiang et al. 2019). Xiang et al. 2019 designed a framework of estimating Functional-And-Evolutionary Trait Heritability scores (FAETH) by combining the information of functionality, evolution, and complex trait heritability, which are used as genomic priors for genomic prediction. This strategy of using priors on functional SNPs in genomic prediction has already proven to be valuable (MacLeod et al. 2016). However, to implement the functional variants for breeding, algorithms for genomic selection should be adapted. To accommodate functional priors in the widely used GBLUP algorithm (Genomic Best Linear Unbiased Predictors Analysis), Fand et al. 2017 designed a GFBLUP algorithm that can take biological priors to improve genomic prediction accuracies (Fang et al. 2017). Integrating genomic selection with functional breeding approaches is now widely explored in both animal and plant breeding, and seems to gain more attention among breeders (Xu et al. 2017).

9.5 Bioinformatics in (animal) breeding

With the adoption of genomic selection, breeding companies have now produced a vast amount of genomics data. Hundreds of thousands of selection candidates have been genotyped, and hundreds of animals have been sequenced (mostly important selection sires). These numbers are increasing steadily. Hence, breeding companies have produced a goldmine of genomics data that can be analysed to identify important variants that alter the phenotype of an individual. However, the fast accumulation of genomics data requires new computational approaches to store and analyse the data, and to implement the findings in the breeding program. These challenges apply to sequence data in particular.

In animal breeding, there has been a long history of research on identifying quantitative trait loci and their application in marker-assisted breeding (Hu et al.

2016). However, the implementation of the findings in breeding programs is still rather limited. This is partly because animal breeding relies on statistical approaches, using prediction models and large numbers, with traditionally less focus on molecular data. In that sense, the quantitative and molecular work seems to come together more closely in recent years (Georges et al. 2019), although the methods to integrate functional information in the breeding programs are still largely under development.

In plant breeding, interestingly, the use of bioinformatics has already been widely adopted (Hu et al. 2018). The development of high-throughput profiling methods for DNA, RNA, and chromatin, complemented by efficient bioinformatics pipelines have greatly enhanced the understanding of genotype to phenotype relationships in plant breeding (Leng et al. 2017). One of the key aspects is to understand the function of important genes, how the genes are regulated (and interact), and which elements in the genome are involved. Traditionally, plant breeders were more interested in molecular genomics work compared to animal breeders. One of the drivers of this interest is to understand disease resistance in plants (Kankanala et al. 2019). Disease resistance is a very important trait in many crop species and gene-editing techniques are now widely applied to improve resistance (Hu et al. 2018). Even disease resistance genes from wild relatives are regularly reintroduced to improve resistance in crops (Dempewolf et al. 2017). To be successful, both methods require a deep understanding of functional elements in the genome that underlie the resistance phenotypes. To that extent, there is a large difference between plant and animal breeders. Though very important, disease resistance traits have generally lower priority in animal breeding programs, and the animals are generally kept in highly confined and sanitized conditions, with lower exposure to pathogens. Moreover, the application of gene-editing is still in its infancy in livestock, partly because of ethical restrictions, and partly because we lack the functional knowledge to identify targets for edits (Tait-Burkard et al. 2018). Nevertheless, with the increase in understanding livestock genomes, aided by the generation of functional data, the interest and application of gene editing will increase. Another important factor why bioinformatics has gained a more prominent role in plant breeding is because many important plant-breeding companies work on multiple species (or within-species strains). This requires a comparative genomics framework to translate information between all these species.

9.6 Future trends

Characterisation of basically all variation present in a breeding line has now become feasible in the breeding industry. While the initial focus was directed towards the identification of deleterious variation, this will now shift to the identification of variation that affects important selection traits in the breeding industry. One key factor to identify high impact variation is to know which regions in the genome are important (i.e. that regulate the expression of genes). This information will be key to understand the relationship between the variant, molecular consequence, and associated phenotype. In addition, the volume of genomic and phenotypic data makes livestock an ideal model for testing genotype-phenotype hypothesis in mammalian species, providing further knowledge of the biological pathways associated with important phenotypes. To that extent, the genome-phenome relationships from livestock populations, and its current resources, will likely become more important to other fields within the life sciences. I believe some domesticated populations are excellent genetic models, because they are outbred, yet with limited effective population size, and with little population stratification. This structure in humans is far more complex, which hampers the efficiency to study the genome-phenome relationship. Finally, I am convinced that a better understanding of the (molecular) mechanisms underlying important selection traits would benefit selection on the long term.

9.7 Concluding remarks

We have entered an exciting era of the ~omics revolution, providing opportunities to bridge the genotype-phenotype gap. In this thesis, I investigated deleterious and functional alleles in various livestock populations. I connect functional genomics, bioinformatics, and breeding data to identify high-impact variation, describing its functional consequences at the molecular, phenotypic, and population level. The results and proposed tools are valuable to infer function from variation, which will be applied to improve livestock breeds in the future.

References

Adams HA, Sonstegard TS, VanRaden PM, Null DJ, Van Tassell CP, Larkin DM, Lewin HA. 2016. Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. *Journal of Dairy Science* **99**: 6693-6701.

Adhikari A, Mainali P, Davie JK. 2019. JARID2 and the PRC2 complex regulates the cell cycle in skeletal muscle. *J Biol Chem* doi:10.1074/jbc.RA119.010060.

Alonso-Spilsbury M, Ramirez-Necoechea R, Gonzalez-Lozano M, Mota-Rojas D, Trujillo-Ortega ME. 2007. Piglet survival in early lactation: A review. *J Anim Vet Adv* **6**: 76-86.

Altshuler DM Durbin RM Abecasis GR Bentley DR Chakravarti A Clark AG Donnelly P Eichler EE Flicek P Gabriel SB et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM (R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789-D798.

Amuzu-Aweh EN, Bovenhuis H, de Koning DJ, Bijma P. 2015. Predicting heterosis for egg production traits in crossbred offspring of individual White Leghorn sires using genome-wide SNP data. *Genetics Selection Evolution* **47**.

Anderson CM, Hu JX, Barnes RM, Heidt AB, Cornelissen I, Black BL. 2015. Myocyte enhancer factor 2C function in skeletal muscle is required for normal growth and glucose metabolism in mice. *Skelet Muscle* **5**.

Andersson L. 2012. How selective sweeps in domestic animals provide new insight into biological mechanisms. *J Intern Med* **271**: 1-14.

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C et al. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* **16**: 57.

Aslam ML, Bastiaansen JWM, Elferink MG, Megens HJ, Crooijmans RPMA, Blomberg L, Fleischer RC, Van Tassell CP, Sonstegard TS, Schroeder SG et al. 2012. Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). *BMC genomics* **13**.

Asmundson V, Abbott UK. 1961. Dominant Sex-Linked Late-Feathering in Turkey. *J Hered* **52**: 99-&.

Ayadi A, Birling MC, Bottomley J, Bussell J, Fuchs H, Fray M, Gailus-Durner V, Greenaway S, Houghton R, Karp N et al. 2012. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian genome : official journal of the International Mammalian Genome Society* **23**: 600-610.

Backstrom N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, Webster MT, Ost T, Schneider M, Kempenaers B et al. 2010. The recombination landscape of the zebra finch Taeniopygia guttata genome. *Genome Research* **20**: 485-495.

Bao Y, Kishnani P, Wu JY, Chen YT. 1996. Hepatic and neuromuscular forms of glycogen storage disease type IV caused by mutations in the same glycogen-branching enzyme gene. *J Clin Invest* **97**: 941-948.

Baribault H, Danao J, Gupte J, Yang L, Sun B, Richards W, Tian H. 2006. The G-protein-coupled receptor GPR103 regulates bone formation. *Molecular and cellular biology* **26**: 709-717.

Barton NH. 2000. Genetic hitchhiking. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **355**: 1553-1562.

Bartonicek N, Clark MB, Quek XC, Torpy JR, Pritchard AL, Maag JLV, Gloss BS, Crawford J, Taft RJ, Hayward NK et al. 2017. Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biol* **18**: 241.

Basson MA. 2012. Signaling in cell differentiation and morphogenesis. *Cold Spring Harb Perspect Biol* **4**.

Baxter EM, Rutherford KMD, D'Eath RB, Arnott G, Turner SP, Sandoe P, Moustsen VA, Thorup F, Edwards SA, Lawrence AB. 2013. The welfare implications of large litter size in the domestic pig II: management factors. *Anim Welfare* **22**: 219-238.

Bianco E, Nevado B, Ramos-Onsins SE, Perez-Enciso M. 2015. A deep catalog of autosomal single nucleotide variation in the pig. *Plos One* **10**: e0118867.

Bioinformatics B. 2011. FastQC: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute*.

Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, the Mouse Genome Database G. 2017. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res* **45**: D723-D729.

Bortoluzzi C, Bosse M, Derks MFL, Crooijmans RPMA, Groenen MAM, Megens HJ. 2019. The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. *Evol Appl* doi:10.1111/eva.12872.

Bosse M, Megens H-J, Derks MFL, de Cara MÁR, Groenen MAM. 2018. Deleterious alleles in the context of domestication, inbreeding and selection. *Evol Appl* **0**.

Bosse M, Megens HJ, Madsen O, Crooijmans RPMA, Ryder OA, Austerlitz F, Groenen MAM, de Cara MAR. 2015. Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research* **25**: 970-981.

Bosse M, Megens HJ, Madsen O, Frantz LA, Paudel Y, Crooijmans RP, Groenen MA. 2014. Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. *Molecular ecology* **23**: 4089-4102.

Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LA, Schook LB, Crooijmans RP, Groenen MA. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *Plos Genet* **8**: e1003100.

Bouquet A, Juga J. 2013. Integrating genomic selection into dairy cattle breeding programmes: a review. *Animal : an international journal of animal bioscience* **7**: 705-713.

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *American journal of human genetics* **103**: 338-348.

Browning SR, Browning BL. 2007. 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* **81**: 1084-1097.

Bu G, Ying Wang C, Cai G, Leung FC, Xu M, Wang H, Huang G, Li J, Wang Y. 2013a. Molecular characterization of prolactin receptor (cPRLR) gene in chickens: gene structure, tissue expression, promoter analysis, and its interaction with chicken prolactin (cPRL) and prolactin-like protein (cPRL-L). *Molecular and cellular endocrinology* **370**: 149-162.

Bu GX, Huang G, Fu H, Li J, Huang SM, Wang YJ. 2013b. Characterization of the novel duplicated PRLR gene at the late-feathering K locus in Lohmann chickens. *J Mol Endocrinol* **51**: 261-276.

Cannon ME, Mohlke KL. 2018. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *American journal of human genetics* **103**: 637-653.

Casas E, Kehrli ME, Jr. 2016. A Review of Selected Genes with Known Effects on Performance and Health of Cattle. *Frontiers in veterinary science* **3**: 113.

Casbon J. 2012. PyVCF - A Variant Call Format Parser for Python.

Casellas J, Canas-Alvarez JJ, Gonzalez-Rodriguez A, Puig-Oliveras A, Fina M, Piedrafita J, Molina A, Diaz C, Baro JA, Varona L. 2017. Bayesian analysis of parent-specific transmission ratio distortion in seven Spanish beef cattle breeds. *Animal Genetics* **48**: 93-96.

Casiro S, Velez-Irizarry D, Ernst CW, Raney NE, Bates RO, Charles MG, Steibel JP. 2017. Genome-wide association study in an F2 Duroc x Pietrain resource population for economically important meat quality and carcass traits. *J Anim Sci* **95**: 545-558.

Cassady JP, Young LD, Leymaster KA. 2002. Heterosis and recombination effects on pig reproductive traits. *J Anim Sci* **80**: 2303-2315.

Castrillo A, Pennington DJ, Otto F, Parker PJ, Owen MJ, Bosca L. 2001. Protein kinase C epsilon is required for macrophage activation and defense against bacterial infection. *J Exp Med* **194**: 1231-1242.

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

Charlesworth B. 2007. A hitch-hiking guide to the genome: a commentary on 'The hitch-hiking effect of a favourable gene' by John Maynard Smith and John Haigh. *Genet Res* **89**: 389-390.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**: 195-205.

Charlesworth D, Willis JH. 2009. FUNDAMENTAL CONCEPTS IN GENETICS The genetics of inbreeding depression. *Nature Reviews Genetics* **10**: 783-796.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F et al. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res* **26**: 1333-1341.

Chen H, Li Z, Haruna K, Li Z, Li Z, Semba K, Araki M, Yamamura K, Araki K. 2008. Early pre-implantation lethality in mice carrying truncated mutation in the RNA polymerase 1-2 gene. *Biochem Biophys Res Commun* **365**: 636-642.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**: 966-968.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**: 2745-2747.

Christianson WT. 1992. Stillbirths, Mummies, Abortions, and Early Embryonic Death. *Vet Clin N Am-Food A* **8**: 623-639.

Chun S, Fay JC. 2011. Evidence for Hitchhiking of Deleterious Mutations within the Human Genome. *Plos Genet* **7**.

Chung J, Zhang X, Collins B, Sper RB, Gleason K, Simpson S, Koh S, Sommer J, Flowers WL, Petters RM et al. 2018. High mobility group A2 (HMGA2) deficiency in pigs leads to dwarfism, abnormal fetal resource allocation, and cryptorchidism. *P Natl Acad Sci USA* **115**: 5420-5425.

Cieslak M, Reissmann M, Hofreiter M, Ludwig A. 2011. Colours of domestication. *Biological reviews of the Cambridge Philosophical Society* **86**: 885-899.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly* **6**: 80-92.

Ciobanu D, Bastiaansen J, Malek M, Helm J, Woollard J, Plastow G, Rothschild M. 2001. Evidence for new alleles in the protein kinase adenosine monophosphate-activated gamma(3)-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. *Genetics* **159**: 1151-1162.

Cole JB, VanRaden PM, Null DJ, Hutchison JL, Cooper TA, Hubbard SM. 2018. Haplotype tests for recessive disorders that affect fertility and other traits. Animal Improvement Program, Animal Genomics and Improvement Laboratory., Agricultural Research Service, USDA, Beltsville.

Craven AJ, Ormandy CJ, Robertson FG, Wilkins RJ, Kelly PA, Nixon AJ, Pearson AJ. 2001. Prolactin signaling influences the timing mechanism of the hair follicle: Analysis of hair growth cycles in prolactin receptor knockout mice. *Endocrinology* **142**: 2533-2539.

Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561-563.

Croft JB, Morrell D, Chase CL, Swift M. 1995. Obesity in heterozygous carriers of the gene for the Bardet-Biedl syndrome. *Am J Med Genet* **55**: 12-15.

Curik I, Ferencakovic M, Solkner J. 2014. Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livest Sci* **166**: 26-34.

Da Silva CLA, Broekhuijse MLWJ, Laurenssen BFA, Mulder HA, Knol EF, Kemp B, Soede NM. 2017. Relationship between ovulation rate and embryonic characteristics in gilts at 35 d of pregnancy. *J Anim Sci* **95**: 3160-3172.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg L, Bouffard P, Burt DW, Crasta O, Crooijmans RPMA et al. 2010. Multi-Platform Next-Generation Sequencing of the Domestic Turkey (Meleagris gallopavo): Genome Assembly and Analysis. *Plos Biol* **8**.

Dalton J, Moore D, Spencer T, Hansen P, Cole J, Neibergs H. 2015. Genomic Selection and Reproductive Efficiency in Dairy Cattle. In *Dairy Cattle Reproduction Council Proceedings*.

Dekkers JC. 2012. Application of genomics tools to animal breeding. *Curr Genomics* **13**: 207-212.

Dempewolf H, Baute G, Anderson J, Kilian B, Smith C, Guarino L. 2017. Past and Future Use of Wild Relatives in Crop Breeding. *Crop Sci* **57**: 1070-1082.

den Hartigh LJ, Wang SR, Goodspeed L, Ding YL, Averill M, Subramanian S, Wietecha T, O'Brien KD, Chait A. 2014. Deletion of Serum Amyloid A3 Improves High Fat High Sucrose Diet-Induced Adipose Tissue Inflammation and Hyperlipidemia in Female Mice. *Plos One* **9**.

Denholm L, Marron, B., Parnell, P., Teseling, C., Beever, J. 2015. Pleiotropic embryonic malformation associated with cranial and caudal neural tube defects from a single amino-acid substitution (V311A) at a conserved locus in the Nhlrc2 protein. *Proceedings of the 9th International Conference on Neural Tube Defects*.

Derks MFL, Gjuvsland AB, Bosse M, Lopes MS, van Son M, Harlizius B, Tan BF, Hamland H, Grindflek E, Groenen MAM et al. 2019. Loss of function mutations in essential genes cause embryonic lethality in pigs. *Plos Genet* **15**.

Derks MFL, Lopes MS, Bosse M, Madsen O, Dibbits B, Harlizius B, Groenen MAM, Megens HJ. 2018. Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. *Plos Genet* **14**: e1007661.

Derks MFL, Megens HJ, Bosse M, Lopes MS, Harlizius B, Groenen MAM. 2017. A systematic survey to identify lethal recessive variation in highly managed pig populations. *BMC genomics* **18**: 858.

Devaux JJ. 2010. The C-terminal domain of beta IV-spectrin is crucial for KCNQ2 aggregation and excitability at nodes of Ranvier. *J Physiol-London* **588**: 4719-4730.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dron N, Hernandez-Jover M, Doyle RE, Holyoake PK. 2014. Investigating risk factors and possible infectious aetiologies of mummified fetuses on a large piggery in Australia. *Australian veterinary journal* **92**: 472-478.

Dumitrache LC, McKinnon PJ. 2017. Polynucleotide kinase-phosphatase (PNKP) mutations and neurologic disease. *Mechanisms of ageing and development* **161**: 121-129.

Elferink MG, Megens HJ, Vereijken A, Hu XX, Crooijmans RPMA, Groenen MAM. 2012. Signatures of Selection in the Genomes of Commercial and Non-Commercial Chicken Breeds. *Plos One* **7**.

Elferink MG, Vallee AAA, Jungerius AP, Crooijmans RPMA, Groenen MAM. 2008. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. *BMC genomics* **9**.

Elferink MG, van As P, Veenendaal T, Crooijmans RPMA, Groenen MAM. 2010. Regional differences in recombination hotspots between two chicken populations. *Bmc Genet* **11**.

Ellen ED, Visscher J, van Arendonk JAM, Bijma P. 2008. Survival of laying hens: Genetic parameters for direct and associative effects in three purebred layer lines. *Poultry Sci* **87**: 233-239.

ENCODE. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Grp MGD. 2015. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* **43**: D726-D736.

Essen-Gustavsson B, Granlund A, Benziane B, Jensen-Waern M, Chibalin AV. 2011. Muscle glycogen resynthesis, signalling and metabolic responses following acute exercise in exercise-trained pigs carrying the PRKAG3 mutation. *Experimental physiology* **96**: 927-937.

FAANG-Consortium. 2020. The regulatory GENomE of SWine and CHicken: functional annotation during development.

Fan W, Du F, Liu X. 2019. TRIM66 confers tumorigenicity of hepatocellular carcinoma cells by regulating GSK-3beta-dependent Wnt/beta-catenin signaling. *Eur J Pharmacol* **850**: 109-117.

Fang LZ, Sahana G, Ma PP, Su GS, Yu Y, Zhang SL, Lund MS, Sorensen P. 2017. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC genomics* **18**.

FAOSTAT. 2017. Food and agriculture organization statistical division (FAOSTAT) of the United Nations. .

Fasquelle C, Sartelet A, Li W, Dive M, Tamma N, Michaux C, Druet T, Huijbers IJ, Isacke CM, Coppieters W et al. 2009. Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. *Plos Genet* **5**: e1000666.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**: 2503-2505.

Freeman TC, Ivens A, Baillie JK, Beraldi D, Barnett MW, Dorward D, Downing A, Fairbairn L, Kapetanovic R, Raza S et al. 2012. A gene expression atlas of the domestic pig. *Bmc Biol* **10**.

Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, Baur A, Grohs C, Weiss B, Boussaha M, Esquerre D et al. 2013. Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. *Plos One* **8**.

Froese DS, Michaeli A, McCorvie TJ, Krojer T, Sasi M, Melaev E, Goldblum A, Zatsepin M, Lossos A, Alvarez R et al. 2015. Structural basis of glycogen branching enzyme deficiency and pharmacologic rescue by rational peptide design. *Human Molecular Genetics* **24**: 5667-5676.

Fujii J, Otsu K, Zorzato F, de Leon S, Khanna VK, Weiler JE, O'Brien PJ, MacLennan DH. 1991. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* **253**: 448-451.

Funari VA, Krakow D, Nevarez L, Chen Z, Funari TL, Vatanavicharn N, Wilcox WR, Rimoin DL, Nelson SF, Cohn DH. 2010. BMPER mutation in diaphanospondylodysostosis identified by ancestral autozygosity mapping and targeted high-throughput sequencing. *American journal of human genetics* **87**: 532-537.

Gallagher MD, Chen-Plotkin AS. 2018. The Post-GWAS Era: From Association to Function. *American journal of human genetics* **102**: 717-730.

Gao TS, He B, Liu S, Zhu H, Tan K, Qian J. 2016. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**: 3543-3551.

Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M. 2015. An estimate of the average number of recessive lethal mutations carried by humans. *Genetics* **199**: 1243-1254.

Garrick DJ, Taylor JF, Fernando RL. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol* **41**: 55.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv* **preprint arXiv:1207.3907 [q-bio.GN]**.

Gautier M, Vitalis R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**: 1176-1177.

Georges M, Charlier C, Hayes B. 2019. Harnessing genomic information for livestock improvement. *Nat Rev Genet* **20**: 135-156.

Gheyas AA, Boschiero C, Eory L, Ralph H, Kuo R, Woolliams JA, Burt DW. 2015. Functional classification of 15 million SNPs detected from diverse chicken populations. *DNA Res* **22**: 205-217.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**: 387-391.

Gilmour AR, Gogel B, Cullis B, Thompson R, Butler D. 2009. ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Ginestet C. 2011. ggplot2: Elegant Graphics for Data Analysis. *J R Stat Soc a Stat* **174**: 245-245.

Giuffra E, Tuggle CK, Consortium F. 2019. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annual review of animal biosciences* **7**: 65-88.

Gjuvsland AB, Vik JO, Beard DA, Hunter PJ, Omholt SW. 2013. Bridging the genotype-phenotype gap: what does it take? *The Journal of physiology* **591**: 2055-2066.

Glemin S. 2003. How are deleterious mutations purged? Drift versus nonrandom mating. *Evolution* **57**: 2678-2687.

Gluecksohn-Waelsch S. 1963. Lethal Genes and Analysis of Differentiation. *Science* **142**: 1269-1276.

Goddard ME, Kemper KE, MacLeod IM, Chamberlain AJ, Hayes BJ. 2016. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *P Roy Soc B-Biol Sci* **283**.

Gonzalez-Pena D, Knox RV, MacNeil MD, Rodriguez-Zas SL. 2015. Genetic gain and economic values of selection strategies including semen traits in three- and four-way crossbreeding systems for swine production. *J Anim Sci* **93**: 879-891.

Grako KA, Ochiya T, Barritt D, Nishiyama A, Stallcup WB. 1999. PDGF alpha-receptor is unresponsive to PDGF-AA in aortic smooth muscle cells from the NG2 knockout mouse. *J Cell Sci* **112**: 905-915.

Groenen MA Archibald AL Uenishi H Tuggle CK Takeuchi Y Rothschild MF Rogel-Gaillard C Park C Milan D Megens HJ et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**: 393-398.

Gross C, de Ridder D, Reinders M. 2018. Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse. *Bmc Bioinformatics* **19**.

Gross C, Derks MFL, Megens HJ, Bosse M, Groenen MA, Reinders M, De Ridder D. 2020. pCADD: SNV prioritisation in Sus scrofa. *Genetics Selection Evolution*.

Gussow AB, Petrovski S, Wang QL, Allen AS, Goldstein DB. 2016. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol* **17**.

Habier D, Fernando RL, Garrick DJ. 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. *Genetics* **194**: 597-+.

Haggman J, Uimari P. 2016. Novel harmful recessive haplotypes for reproductive traits in pigs. *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* doi:10.1111/jbg.12240.

Hall SJ. 2016. Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. *Animal : an international journal of animal bioscience* **10**: 1778-1785.

Halpain S, Dehmelt L. 2006. The MAP1 family of microtubule-associated proteins. *Genome Biol* **7**: 224.

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, Mullaart E, Coppieters W, Georges M. 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. *bioRxiv* doi:10.1101/079863.

Hayes B, Goddard ME. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol* **33**: 209-229.

Hedrick PW. 2015. Heterozygote advantage: the effect of artificial selection in livestock and pets. *J Hered* **106**: 141-154.

Hedrick PW, Garcia-Dorado A. 2016. Understanding Inbreeding Depression, Purging, and Genetic Rescue. *Trends in ecology & evolution* **31**: 940-952.

Helbing T, Rothweiler R, Ketterer E, Goetz L, Heinke J, Grundmann S, Duerschmied D, Patterson C, Bode C, Moser M. 2011. BMP activity controlled by BMPER regulates the proinflammatory phenotype of endothelium. *Blood* **118**: 5040-5049.

Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. 2017. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet* **95**: 1 22 21-21 22 23.

Hernandez SC, Finlayson HA, Ashworth CJ, Haley CS, Archibald AL. 2014. A genome-wide linkage analysis for reproductive traits in F2 Large White x Meishan cross gilts. *Anim Genet* **45**: 191-197.

Hidalgo AM, Bastiaansen JWM, Lopes MS, Calus MPL, de Koning DJ. 2016. Accuracy of genomic prediction of purebreds for cross bred performance in pigs. *Journal of Animal Breeding and Genetics* **133**: 443-451.

Hillier LW Miller W Birney E Warren W Hardison RC Ponting CP Bork P Burt DW Groenen MAM Delany ME et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.

Hoff JL, Decker JE, Schnabel RD, Taylor JF. 2017. Candidate lethal haplotypes and causal mutations in Angus cattle. *BMC genomics* **18**: 799.

Hong J, Kim D, Cho K, Sa S, Choi S, Kim Y, Park J, Schmidt GS, Davis ME, Chung H. 2015. Effects of genetic variants for the swine FABP3, HMGA1, MC4R, IGF2, and FABP4 genes on fatty acid composition. *Meat science* **110**: 46-51.

Hoving R, Hulsegge I, Hiemstra S. 2017. Varkenrassen in de genenbank. *Centrum voor Genetische Bronnen, Nederland (CGN) van Wageningen University & Research* **CGN rapport 37**.

Howard DM, Pong-Wong R, Knap PW, Woolliams JA. 2017a. Use of haplotypes to identify regions harbouring lethal recessive variants in pigs. *Genetics Selection Evolution* **49**.

Howard JT, Pryce JE, Baes C, Maltecca C. 2017b. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. *J Dairy Sci* **100**: 6009-6024.

Hu HF, Scheben A, Edwards D. 2018. Advances in Integrating Genomics and Bioinformatics in the Plant Breeding Pipeline. *Agriculture-Basel* **8**.

Hu ZL, Park CA, Reecy JM. 2016. Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res* **44**: D827-833.

Hu ZL, Park CA, Reecy JM. 2019. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res* **47**: D701-D710.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* **8**.

Hug N, Longman D, Caceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res* **44**: 1483-1495.

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P et al. 2018. Ensembl variation resources. *Database-Oxford* doi:ARTN bay119

10.1093/database/bay119.

Ikeya M, Kawada M, Kiyonari H, Sasai N, Nakao K, Furuta Y, Sasai Y. 2006. Essential pro-Bmp roles of crossveinless 2 in mouse organogenesis. *Development* **133**: 4463-4473.

Imsland F, Feng CG, Boije H, Bed'hom B, Fillon V, Dorshorst B, Rubin CJ, Liu RR, Gao Y, Gu XR et al. 2012. The Rose-comb Mutation in Chickens Constitutes a Structural Rearrangement Causing Both Altered Comb Morphology and Defective Sperm Motility. *Plos Genet* **8**.

Iversen MW, Nordbo O, Gjerlaug-Enger E, Grindflek E, Lopes MS, Meuwissen THE. 2017. Including crossbred pigs in the genomic relationship matrix through utilization of both linkage disequilibrium and linkage analysis. *J Anim Sci* **95**: 5197-5207.

Jang MJ, Park UH, Kim JW, Choi H, Um SJ, Kim EJ. 2017. CACUL1 reciprocally regulates SIRT1 and LSD1 to repress PPARgamma and inhibit adipogenesis. *Cell death & disease* **8**: 3201.

Jiang Y, Ding Q, Xie XL, Libby RT, Lefebvre V, Gan L. 2013. Transcription Factors SOX4 and SOX11 Function Redundantly to Regulate the Development of Mouse Retinal Ganglion Cells. *J Biol Chem* **288**: 18429-18438.

Jimenez-Mena B, Hospital F, Bataillon T. 2016. Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conserv Genet Resour* **8**: 35-41.

Johnson PA, Stephens CS, Giles JR. 2015. The domestic chicken: Causes and consequences of an egg a day. *Poult Sci* **94**: 816-820.

Joshi NA, Fass JN. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33) [Software].

Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, Karim L, Nielsen US, Panitz F, Aamand GP, Schulman N et al. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *Plos Genet* **10**: e1004049.

Kallabi F, Hadj Salem I, Ben Chehida A, Ben Salah G, Ben Turkia H, Tebib N, Keskes L, Kamoun H. 2015. Splicing defects in ABCD1 gene leading to both exon skipping and partial intron retention in X-linked adrenoleukodystrophy Tunisian patient. *Neurosci Res* **97**: 7-12.

Kankanala P, Nandety RS, Mysore KS. 2019. Genomics of Plant Disease Resistance in Legumes. *Frontiers in plant science* **10**: 1345.

Kearney JF, Navarro P, Haley CS, Villanueva B. 2009. Consequences of selection for improving production traits on the frequency of deleterious alleles for fitness. *J Anim Sci* **87**: 850-859.

Kelleher MM, Berry DP, Kearney JF, McParland S, Buckley F, Purfield DC. 2017. Inference of population structure of purebred dairy and beef cattle using high-density genotype data. *Animal : an international journal of animal bioscience* **11**: 15-23.

Kelley R, Ren R, Pi X, Wu Y, Moreno I, Willis M, Moser M, Ross M, Podkowa M, Attisano L et al. 2009. A concentration-dependent endocytic trap and sink mechanism converts Bmper from an activator to an inhibitor of Bmp signaling. *The Journal of cell biology* **184**: 597-609.

Khan MA, Mohan S, Zubair M, Windpassinger C. 2016. Homozygosity mapping identified a novel protein truncating mutation (p.Ser100Leufs*24) of the BBS9 gene in a consanguineous Pakistani family with Bardet Biedl syndrome. *BMC medical genetics* **17**: 10.

Kim KS, Larsen N, Short T, Plastow G, Rothschild MF. 2000. A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. *Mammalian genome : official journal of the International Mammalian Genome Society* **11**: 131-135.

Kim KS, Lee JJ, Shin HY, Choi BH, Lee CK, Kim JJ, Cho BW, Kim TH. 2006. Association of melanocortin 4 receptor (MC4R) and high mobility group AT-hook 1 (HMGA1) polymorphisms with pig growth and fat deposition traits. *Anim Genet* **37**: 419-421.

Klink BU, Zent E, Juneja P, Kuhlee A, Raunser S, Wittinghofer A. 2017. A recombinant BBSome core complex and how it interacts with ciliary cargo. *Elife* **6**.

Knierim E, Gill E, Seifert F, Morales-Gonzalez S, Unudurthi SD, Hund TJ, Stenzel W, Schuelke M. 2017. A recessive mutation in beta-IV-spectrin (SPTBN4) associates with congenital myopathy, neuropathy, and central deafness. *Human genetics* **136**: 903-910.

Knol EF, Ducro BJ, van Arendonk JAM, van der Lende T. 2002. Direct, maternal and nurse sow genetic effects on farrowing-, pre-weaning- and total piglet survival. *Livest Prod Sci* **73**: 153-164.

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

Komada M, Soriano P. 2002. beta IV-spectrin regulates sodium channel clustering through ankyrin-G at axon initial segments and nodes of Ranvier. *Journal of Cell Biology* **156**: 337-348.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471-475.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639-1645.

Kulaga HM, Leitch CC, Eichers ER, Badano JL, Lesemann A, Hoskins BE, Lupski JR, Beales PL, Reed RR, Katsanis N. 2004. Loss of BBS proteins causes anosmia in humans and defects in olfactory cilia structure and function in the mouse. *Nat Genet* **36**: 994-998.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1082.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Le Cozler Y, Guyomarc'h C, Pichodo X, Quinio PY, Pellois H. 2002. Factors associated with stillborn and mummified piglets in high-prolific sows. *Anim Res* **51**: 261-268.

Lehner B. 2013. Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet* **14**: 168-178.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Leng PF, Lubberstedt T, Xu ML. 2017. Genomics-assisted breeding - A revolutionary strategy for crop improvement. *J Integr Agr* **16**: 2674-2685.

Lentz J, Keats BJB. 1993. Usher Syndrome Type I. In *GeneReviews((R)),* (ed. MP Adam, et al.), Seattle (WA).

Leroy G. 2014. Inbreeding depression in livestock species: review and meta-analysis. *Anim Genet* **45**: 618-628.

Lewandowska MA. 2013. The missing puzzle piece: splicing mutations. *Int J Clin Exp Patho* **6**: 2675-2682.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li X, Kim SW, Choi JS, Lee YM, Lee CK, Choi BH, Kim TH, Choi YI, Kim JJ, Kim KS. 2010. Investigation of porcine FABP3 and LEPR gene polymorphisms and mRNA expression for variation in intramuscular fat content. *Mol Biol Rep* **37**: 3931-3939.

Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, Sherlock RG, Li WB, Lukefahr SD, Shanks BC et al. 2014. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nat Commun* **5**.

Liu YH, Qi WB, Richardson A, Van Remmen H, Ikeno Y, Salmon AB. 2013. Oxidative damage associated with obesity is prevented by overexpression of CuZn- or Mn-superoxide dismutase. *Biochem Bioph Res Co* **438**: 78-83.

Liu Z, Sun WX, Zhao YY, Xu CY, Fu YY, Li Y, Chen J. 2014. The effect of variants in the promoter of BMPER on the intramuscular fat deposition in longissimus dorsi muscle of pigs. *Gene* **542**: 168-172.

Lu Z, Yuan Z, Miyoshi T, Wang Q, Su Z, Chang CC, Shi W. 2011. Identification of Soat1 as a quantitative trait locus gene on mouse chromosome 1 contributing to hyperlipidemia. *Plos One* **6**: e25344.

Luo CL, Shen X, Rao YS, Xu HP, Tang J, Sun L, Nie QH, Zhang XQ. 2012. Differences of Z chromosome and genomic expression between early- and late-feathering chickens. *Mol Biol Rep* **39**: 6283-6288.

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC genomics* **17**.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* doi:10.1093/nar/gkz268.

Mahlapuu M, Johansson C, Lindgren K, Hjalm G, Barnes BR, Krook A, Zierath JR, Andersson L, Marklund S. 2004. Expression profiling of the gamma-subunit isoforms of AMP-activated protein kinase suggests a major role for gamma3 in white skeletal muscle. *American journal of physiology Endocrinology and metabolism* **286**: E194-200.

Makino T, Rubin CJ, Carneiro M, Axelsson E, Andersson L, Webster MT. 2018. Elevated Proportions of Deleterious Genetic Variation in Domestic Animals and Plants. *Genome Biol Evol* **10**: 276-290.

Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vila C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *P Natl Acad Sci USA* **113**: 152-157.

Matika O, Robledo D, Pong-Wong R, Bishop SC, Riggio V, Finlayson H, Lowe NR, Hoste AE, Walling GA, del Pozo J et al. 2019. Balancing selection at a premature stop mutation in the myostatin gene underlies a recessive leg weakness syndrome in pigs. *Plos Genet* **15**.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010a. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297-1303.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010b. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**: 1297-1303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122.

Megens HJ, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, Vereijken A, Silva P, Muir WM, Cheng HH et al. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. *Bmc Genet* **10**.

Merks J. 2000. One century of genetic changes in pigs and the future needs. *BSAP Occasional Publication* **27**: 8-19.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Milan D, Jeon JT, Looft C, Amarger V, Robic A, Thelander M, Rogel-Gaillard C, Paul S, Iannuccelli N, Rask L et al. 2000. A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288**: 1248-1251.

Moser M, Binder O, Wu Y, Aitsebaomo J, Ren R, Bode C, Bautch VL, Conlon FL, Patterson C. 2003. BMPER, a novel endothelial cell precursor-derived protein, antagonizes bone morphogenetic protein signaling and endothelial cell differentiation. *Molecular and cellular biology* **23**: 5664-5679.

Mu KD, Sun Y, Zhao Y, Zhao TX, Li Q, Zhang ML, Li HT, Zhang R, Hu C, Wang C et al. 2019. Hepatic nitric oxide synthase 1 adaptor protein regulates glucose homeostasis and hepatic insulin sensitivity in obese mice depending on its PDZ binding domain. *Ebiomedicine* **47**: 352-364.

Mulder HA, Lee SH, Clark S, Hayes BJ, van der Werf JHJ. 2019. The Impact of Genomic and Traditional Selection on the Contribution of Mutational Variance to Long-Term Selection Response and Genetic Variance. *Genetics* **213**: 361-378.

Murgiano L, Tammen I, Harlizius B, Drogemuller C. 2012. A de novo germline mutation in MYH7 causes a progressive dominant myopathy in pigs. *Bmc Genet* **13**.

Nagy L, Marton J, Vida A, Kis G, Bokor E, Kun S, Gonczi M, Docsa T, Toth A, Antal M et al. 2018. Glycogen phosphorylase inhibition improves beta cell function. *Br J Pharmacol* **175**: 301-319.

Nakamura A, Ishikawa A, Nagao K, Watanabe H, Uchida M, Kansaku N. 2011. Characteristics of Reversion to Early Feathering Phenotype in the Late Feathering Line of Nagoya Breed Chickens. *J Poult Sci* **48**: 155-161.

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic Variation in an Individual Human Exome. *Plos Genet* **4**.

Nicholas FW. 2003. Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic Acids Res* **31**: 275-277.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443-451.

Nishimura DY, Swiderski RE, Searby CC, Berg EM, Ferguson AL, Hennekam R, Merin S, Weleber RG, Biesecker LG, Stone EM et al. 2005. Comparative genomics and gene expression analysis identifies BBS9, a new Bardet-Biedl syndrome gene. *American journal of human genetics* **77**: 1021-1033.

Niwa M, Numaguchi Y, Ishii M, Kuwahata T, Kondo M, Shibata R, Miyata K, Oike Y, Murohara T. 2015. IRAP deficiency attenuates diet-induced obesity in mice through increased energy expenditure. *Biochem Biophys Res Commun* **457**: 12-18.

Novas R, Cardenas-Rodriguez M, Irigoin F, Badano JL. 2015. Bardet-Biedl syndrome: Is it only cilia dysfunction? *FEBS Lett* **589**: 3479-3491.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I et al. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *Plos Genet* **10**: e1004234.

Okonechnikov K, Conesa A, Garcia-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**: 292-294.

Olijslagers H. 2018. Balanced Breeding Pays Off. Topigs Norsvin.

Onteru SK, Fan B, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF. 2012. A whole-genome association study for pig reproductive traits. *Anim Genet* **43**: 18-26.

Ormandy CJ, Camus A, Barra J, Damotte D, Lucas B, Buteau H, Edery M, Brousse N, Babinet C, Binart N et al. 1997. Null mutation of the prolactin receptor gene produces multiple reproductive defects in the mouse. *Genes & development* **11**: 167-178.

Palmiter RD. 2009. Reduced levels of neurotransmitter-degrading enzyme PRCP promote obesity. *Journal of Clinical Investigation* **119**: 2130-2133.

Pankotai T, Komonyi O, Bodai L, Ujfaludi Z, Muratoglu S, Ciurciu A, Tora L, Szabad J, Boros I. 2005. The homologous Drosophila transcriptional adaptors ADA2a and ADA2b are both required for normal development but have different functions. *Molecular and cellular biology* **25**: 8215-8227.

Parkinson NJ, Olsson CL, Hallows JL, McKee-Johnson J, Keogh BP, Noben-Trauth K, Kujawa SG, Tempel BL. 2001. Mutant beta-spectrin 4 causes auditory and motor neuropathies in quivering mice. *Nat Genet* **29**: 61-65.

Paudel Y, Madsen O, Megens HJ, Frantz LA, Bosse M, Crooijmans RP, Groenen MA. 2015. Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. *BMC genomics* **16**: 330.

Pausch H, Schwarzenbacher H, Burgstaller J, Flisikowski K, Wurmser C, Jansen S, Jung S, Schnieke A, Wittek T, Fries R. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. *BMC genomics* **16**: 312.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**: 2825-2830.

Perez-Enciso M, Forneris N, de los Campos G, Legarra A. 2017. Evaluating Sequence-Based Genomic Prediction with an Efficient New Simulator. *Genetics* **205**: 939-953.

Perez-Enciso M, Rincon JC, Legarra A. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genetics Selection Evolution* **47**.

Perleberg C, Kind A, Schnieke A. 2018. Genetically engineered pigs as models for human disease. *Dis Model Mech* **11**.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *Plos Genet* **9**: e1003709.

Pinhasov A, Mandel S, Torchinsky A, Giladi E, Pittel Z, Goldsweig AM, Servoss SJ, Brenneman DE, Gozes I. 2003. Activity-dependent neuroprotective protein: a novel gene essential for brain formation. *Dev Brain Res* **144**: 83-90.

Piorkowska K, Ropka-Molik K, Szmatola T, Zygmunt K, Tyra M. 2014. Association of a new mobile element in predicted promoter region of ATP-binding cassette transporter 12 gene (ABCA12) with pig production traits. *Livest Sci* **168**: 38-44.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Research* **21**: 1769-1776.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**: 559-575.

Qian CM, Zhang Q, Li SD, Zeng L, Walsh MJ, Zhou MM. 2005. Structure and chromosomal DNA binding of the SWIRM domain. *Nat Struct Mol Biol* **12**: 1078-1085.

Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi- C, Bredeson JV, Bart RS, Verma J, Buckler ES, Lu F. 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat Genet* **49**: 959-+.

Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. 2018a. Utility of whole-genome sequence data for across-breed genomic prediction. *Genetics Selection Evolution* **50**.

Raymond B, Bouwman AC, Wientjes YCJ, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. 2018b. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. *Genetics Selection Evolution* **50**.

Renema RA, Sikur VR, Robinson FE, Korver DR, Zuidhof MJ. 2008. Effects of nutrient density and age at photostimulation on carcass traits and reproductive efficiency in fast- and slow-feathering turkey hens. *Poultry Sci* **87**: 1897-1908.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**: D886-D894.

Reyer H, Ponsuksili S, Wimmers K, Murani E. 2013. Transcript variants of the porcine glucocorticoid receptor gene (NR3C1). *Gen Comp Endocr* **189**: 127-133.

Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* **42**: W320-324.

Rojano E, Seoane P, Ranea JAG, Perkins JR. 2019. Regulatory variants: from detection to predicting impact. *Brief Bioinform* **20**: 1639-1654.

Ron M, Weller JI. 2007. From QTL to QTN identification in livestock - winning by points rather than knock-out: a review. *Animal Genetics* **38**: 429-439.

Rosenvold K, Petersen JS, Laerke HN, Jensen SK, Therkildsen M, Karlsson AH, Moller HS, Andersen HJ. 2001. Muscle glycogen stores and meat quality as affected by strategic finishing feeding of slaughter pigs. *J Anim Sci* **79**: 382-391.

Rothammer S, Kremer PV, Bernau M, Fernandez-Figares I, Pfister-Schar J, Medugorac I, Scholz AM. 2014. Genome-wide QTL mapping of nine body composition and bone mineral density traits in pigs. *Genet Sel Evol* **46**: 68.

Rubin CJ, Megens HJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg O, Jern P, Jorgensen CB et al. 2012. Strong signatures of selection in the domestic pig genome. *P Natl Acad Sci USA* **109**: 19529-19536.

Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587-U145.

Rutherford KMD, Baxter EM, D'Eath RB, Turner SP, Arnott G, Roehe R, Ask B, Sandoe P, Moustsen VA, Thorup F et al. 2013. The welfare implications of large litter size in the domestic pig I: biological factors. *Anim Welfare* **22**: 199-218.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832-837.

Sahana G, Iso-Touru T, Wu X, Nielsen US, de Koning DJ, Lund MS, Vilkki J, Guldbrandtsen B. 2016. A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. *Genet Sel Evol* **48**: 35.

Sahana G, Nielsen US, Aamand GP, Lund MS, Guldbrandtsen B. 2013. Novel harmful recessive haplotypes identified for fertility traits in Nordic Holstein cattle. *Plos One* **8**: e82909.

Savory CJ. 1995. Feather Pecking and Cannibalism. *World Poultry Sci J* **51**: 215-219.

Schaid DJ, Chen W, Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**: 491-504.

Schmid M Smith J Burt DW Aken BL Antin PB Archibald AL Ashwell C Blackshear PJ Boschiero C Brown CT et al. 2015. Third Report on Chicken Genes and Chromosomes 2015. *Cytogenet Genome Res* **145**: 78-179.

Schneider JF, Rempel LA, Snelling WM, Wiedmann RT, Nonneman DJ, Rohrer GA. 2012. Genome-wide association study of swine farrowing traits. Part II: Bayesian analysis of marker data. *J Anim Sci* **90**: 3360-3367.

Schubert M, Jonsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, Albrechtsen A, Dupanloup I, Foucal A, Petersen B et al. 2014. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc Natl Acad Sci U S A* **111**: E5661-5669.

Schutz E, Wehrhahn C, Wanjek M, Bortfeld R, Wemheuer WE, Beck J, Brenig B. 2016. The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. *Plos One* **11**: e0154602.

Shaheen R, Szymanska K, Basu B, Patel N, Ewida N, Faqeih E, Al Hashem A, Derar N, Alsharif H, Aldahmesh MA et al. 2016. Characterizing the morbid genome of ciliopathies. *Genome Biol* **17**.

Sharma A, Lee JS, Dang CG, Sudrajad P, Kim HC, Yeon SH, Kang HS, Lee SH. 2015. Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. *Asian Austral J Anim* **28**: 1371-1379.

Sheffield NC, Thurman RE, Song LY, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research* **23**: 777-788.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. *Nature* **550**: 345-353.

Shimada M, Dumitrache LC, Russell HR, McKinnon PJ. 2015. Polynucleotide kinase-phosphatase enables neurogenesis via multiple DNA repair pathways to maintain genome stability. *EMBO J* **34**: 2465-2480.

Siegel PB, Mueller CD, Craig JV. 1957. Some Phenotypic Differences among Homozygous, Heterozygous, and Hemizygous Late Feathering Chicks. *Poultry Sci* **36**: 232-239.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7**: 539.

Sikur VR, Robinson FE, Korver DR, Renema RA, Zuidhoft MJ. 2004. Effects of nutrient density on growth and carcass traits in fast- and slow-feathering female turkeys. *Poultry Sci* **83**: 1507-1517.

Sinal CJ, Tohkin M, Miyata M, Ward JM, Lambert G, Gonzalez FJ. 2000. Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis. *Cell* **102**: 731-744.

Sironen A, Uimari P, Iso-Touru T, Vilkki J. 2012. L1 insertion within SPEF2 gene is associated with increased litter size in the Finnish Yorkshire population. *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* **129**: 92-97.

Sitzenstock F, Ytournel F, Sharifi AR, Cavero D, Taubert H, Preisinger R, Simianer H. 2013. Efficiency of genomic selection in an established commercial layer breeding program. *Genetics Selection Evolution* **45**.

Sjolund J, Pelorosso FG, Quigley DA, DelRosario R, Balmain A. 2014. Identification of Hipk2 as an essential regulator of white fat development. *P Natl Acad Sci USA* **111**: 7373-7378.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: A next-generation genome browser. *Genome Research* **19**: 1630-1638.

Smeds L, Qvarnstrom A, Ellegren H. 2016. Direct estimate of the rate of germline mutation in a bird. *Genome Research* **26**: 1211-1218.

Sohail M, Vakhrusheva OA, Sul JH, Pulit SL, Francioli LC, Genome of the Netherlands C, Alzheimer's Disease Neuroimaging I, van den Berg LH, Veldink JH, de Bakker PIW et al. 2017. Negative selection in humans and fruit flies involves synergistic epistasis. *Science* **356**: 539-542.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Recombination: the good, the bad and the variable. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **372**.

Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zollner S, Rosenberg NA, Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. *American journal of human genetics* **93**: 90-102.

Tait-Burkard C, Doeschl-Wilson A, McGrew MJ, Archibald AL, Sang HM, Houston RD, Whitelaw CB, Watson M. 2018. Livestock 2.0-genome editing for fitter, healthier, and more productive farmed animals. *Genome Biol* **19**.

Tanabe A, Yanagiya T, Iida A, Saito S, Sekine A, Takahashi A, Nakamura T, Tsunoda T, Kamohara S, Nakata Y et al. 2007. Functional single-nucleotide polymorphisms in the secretogranin III (SCG3) gene that form secretory granules with appetite-related neuropeptides are associated with obesity. *The Journal of clinical endocrinology and metabolism* **92**: 1145-1154.

The UniProt C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**: D158-D169.

Tier B, Meyer K. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models *Journal of Animal Breeding and Genetics* **121**: 77-89.

Tortereau F, Servin B, Frantz L, Megens HJ, Milan D, Rohrer G, Wiedmann R, Beever J, Archibald AL, Schook LB et al. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC genomics* **13**.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46-+.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562-578.

Trask AE, Bignal EM, McCracken DI, Monaghan P, Piertney SB, Reid JM. 2016. Evidence of the phenotypic expression of a lethal recessive allele under inbreeding in a wild population of conservation concern. *The Journal of animal ecology* **85**: 879-891.

Trevaskis J, Walder K, Foletta V, Kerr-Bayles L, McMillan J, Cooper A, Lee S, Bolton K, Prior M, Fahey R et al. 2005. Src homology 3-domain growth factor receptor-bound 2-like (endophilin) interacting protein 1, a novel neuronal protein that regulates energy balance. *Endocrinology* **146**: 3757-3764.

Turner SD. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*.

Uimari P, Sironen A. 2014. A combination of two variants in PRKAG3 is needed for a positive effect on meat quality in pigs. *Bmc Genet* **15**.

UniProt C. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**: D506-D515.

Upperman LR, Kinghorn BP, MacNeil MD, Van Eenennaam AL. 2019. Management of lethal recessive alleles in beef cattle through the use of mate selection software. *Genetics Selection Evolution* **51**.

van Binsbergen R, Calus MPL, Bink MCAM, van Eeuwijk FA, Schrooten C, Veerkamp RF. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* **47**.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**: 1061-1063.

van der Lende T, Knol EF, Leenhouwers JI. 2001. Prenatal development as a predisposing factor for perinatal losses in pigs. *Reprod Suppl* **58**: 247-261.

van der Lende T, van Rens BT. 2003. Critical periods for foetal mortality in gilts identified by analysing the length distribution of mummified foetuses and frequency of non-fresh stillborn piglets. *Animal reproduction science* **75**: 141-150.

van Son M, Kent MP, Grove H, Agarwal R, Hamland H, Lien S, Grindflek E. 2017a. Fine mapping of a QTL affecting levels of skatole on pig chromosome 7. *Bmc Genet* **18**: 85.

van Son M, Lopes MS, Martell HJ, Derks MFL, Gangsei LE, Kongsro J, Wass MN, Grindflek EH, Harlizius B. 2019. A QTL for Number of Teats Shows Breed Specific Effects on Number of Vertebrae in Pigs: Bridging the Gap Between Molecular and Quantitative Genetics. *Frontiers in genetics* **10**.

van Son M, Tremoen NH, Gaustad AH, Myromslien FD, Vage DI, Stenseth EB, Zeremichael TT, Grindflek E. 2017b. RNA sequencing reveals candidate genes and polymorphisms related to sperm DNA integrity in testis tissue from boars. *BMC veterinary research* **13**: 362.

VanRaden PM, Olson KM, Null DJ, Hutchison JL. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *J Dairy Sci* **94**: 6153-6161.

Verardo LL, Silva FF, Lopes MS, Madsen O, Bastiaansen JW, Knol EF, Kelly M, Varona L, Lopes PS, Guimaraes SE. 2016. Revealing new candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways. *Genet Sel Evol* **48**: 9.

Veroneze R, Lopes PS, Guimaraes SEF, Silva FF, Lopes MS, Harlizius B, Knol EF. 2013. Linkage disequilibrium and haplotype block structure in six commercial pig lines. *J Anim Sci* **91**: 3493-3501.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160**: 554-566.

Wang CC, Ortiz-Gonzalez XR, Yum SW, Gill SM, White A, Kelter E, Seaver LH, Lee S, Wiley G, Gaffney PM et al. 2018. betaIV Spectrinopathies Cause Profound Intellectual Disability, Congenital Hypotonia, and Motor Axonal Neuropathy. *American journal of human genetics* **102**: 1158-1168.

Wang GD, Xie HB, Peng MS, Irwin D, Zhang YP. 2014. Domestication genomics: evidence from animals. *Annual review of animal biosciences* **2**: 65-84.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**: 1665-1674.

Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, Chow W, Eory L, Finlayson HA, Flicek P et al. 2019. An improved pig reference genome sequence to enable pig genetics and genomics research. *bioRxiv* doi:10.1101/668921: 668921.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F et al. 2017. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3-Genes Genom Genet* **7**: 109-117.

Watson JD. 2014. *Molecular biology of the gene*. Pearson, Boston.

Webster AB. 2004. Welfare implications of avian osteoporosis. *Poultry Sci* **83**: 184-192.

Willet CE, Wade CM. 2014. From the phenotype to the genotype via bioinformatics. *Methods Mol Biol* **1168**: 1-16.

Wright D. 2015. The Genetic Architecture of Domestication in Animals. *Bioinformatics and biology insights* **9**: 11-20.

Wright S. 1990. Evolution in Mendelian Populations (Reprinted from Genetics, Vol 16, Pg 97-159, 1931). *B Math Biol* **52**: 241-295.

Xiang R, Berg IVD, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, Bolormaa S, Liu Z, Rochfort SJ, Reich CM et al. 2019. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc Natl Acad Sci U S A* **116**: 19398-19408.

Xiang RD, MacLeod IM, Bolormaa S, Goddard ME. 2017. Genome-wide comparative analyses of correlated and uncorrelated phenotypes identify major pleiotropic variants in dairy cattle. *Sci Rep-Uk* **7**.

Xie C, Tammi MT. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *Bmc Bioinformatics* **10**.

Xu YB, Li P, Zou C, Lu YL, Xie CX, Zhang XC, Prasanna BM, Olsen MS. 2017. Enhancing genetic gain in the era of molecular breeding. *J Exp Bot* **68**: 2641-2666.

Yagyu H, Kitamine T, Osuga J, Tozawa R, Chen Z, Kaji Y, Oka T, Perrey S, Tamura Y, Ohashi K et al. 2000. Absence of ACAT-1 attenuates atherosclerosis but causes dry eye and cutaneous xanthomatosis in mice with congenital hyperlipidemia. *J Biol Chem* **275**: 21324-21330.

Yang YL, Zhou R, Li K. 2017. Future livestock breeding: Precision breeding based on multiomics information and population personalization. *J Integr Agr* **16**: 2784-2791.

Yin T, Wensch-Dorendorf M, Simianer H, Swalve HH, Konig S. 2014. Assessing the impact of natural service bulls and genotype by environment interactions on genetic gain and inbreeding in organic dairy cattle genomic breeding programs. *Animal : an international journal of animal bioscience* **8**: 877-886.

Yoshizumi S, Suzuki S, Hirai M, Hinokio Y, Yamada T, Yamada T, Tsunoda U, Aburatani H, Yamaguchi K, Miyagi T et al. 2007. Increased hepatic expression of ganglioside-specific sialidase, NEU3, improves insulin sensitivity and glucose tolerance in mice. *Metabolism: clinical and experimental* **56**: 420-429.

Yun JY, Jin HG, Cao Y, Zhang LC, Zhao YM, Jin X, Yu YS. 2018. RNA-Seq Analysis Reveals a Positive Role of HTR2A in Adipogenesis in Yan Yellow Cattle. *Int J Mol Sci* **19**.

Zak LJ, Gaustad AH, Bolarin A, Broekhuijse M, Walling GA, Knol EF. 2017. Genetic control of complex traits, with a focus on reproduction in pigs. *Mol Reprod Dev* **84**: 1004-1011.

Zakrzewska EI, Savage TF. 1997. Inhibited feathering: A new dominant sex-linked gene in the turkey. *J Hered* **88**: 238-247.

Zappala Z, Montgomery SB. 2016. Non-Coding Loss-of-Function Variation in Human Genomes. *Hum Hered* **81**: 78-87.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754-D761.

Zhang B, Shang P, Qiangba YZ, Xu AS, Wang ZX, Zhang H. 2016a. The association of NR1H3 gene with lipid deposition in the pig. *Lipids Health Dis* **15**.

Zhang H, Yin L, Wang M, Yuan X, Liu X. 2019. Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. *Frontiers in genetics* **10**: 189.

Zhang M, Zhou L, Bawa R, Suren H, Holliday JA. 2016b. Recombination Rate Variation, Hitchhiking, and Demographic History Shape Deleterious Load in Poplar. *Mol Biol Evol* **33**: 2899-2910.

Zhang Q, Seo S, Bugge K, Stone EM, Sheffield VC. 2012. BBS proteins interact genetically with the IFT pathway to influence SHH-related phenotypes. *Hum Mol Genet* **21**: 1945-1953.

Zhao CP, Gui LS, Li YK, Plath M, Zan LS. 2015. Associations between allelic polymorphism of the BMP Binding Endothelial Regulator and phenotypic variation of cattle. *Mol Cell Probe* **29**: 358-364.

Zhao J, Yao J, Li F, Yang Z, Sun Z, Qu L, Wang K, Su Y, Zhang A, Montgomery SA et al. 2016. Identification of candidate genes for chicken early- and late-feathering. *Poultry Sci* **95**: 1498-1503.

Summary

Summary

The DNA provides a blueprint of life containing the instruction, together with the environment, that determine the phenotype. In this thesis I attempt to further close the genotype phenotype gap in livestock, contributing to our understanding of important variation in the animals genomes. I analysed hundred thousands of genotypes, and hundreds of whole genome sequenced individuals to identify variation with impact, either deleterious (e.g. recessive lethals) or variants with positive effects on important selection traits. With this thesis I provide a comprehensive overview high-impact variation in various livestock breeds, and discuss the implications for breeding.

In chapter 2 I perform a survey to assess deleterious haplotypes that likely harbor a recessive lethal allele in three pig populations. I demonstrate that the approach to identify recessive lethal haplotypes from regions that show a deficit in homozygosity can successfully be applied in pigs. Moreover, I report one haplotype that results in relatively late termination of fetal development leading to mummification of homozygous animals.

Chapter 3 describes a sequence-based GWAS study to investigate the molecular basis of the sex-linked feathering rate at hatch in domestic turkey. I describe a 5-bp frameshift deletion in the prolactin receptor (*PRLR*) gene that is responsible for slow feathering at hatch. The consequence, a partial C-terminal loss of the prolactin receptor, is strikingly similar to the protein encoded by the slow feathering K allele in chicken, but with a different causative mutation.

In **chapter 4** I examined the genome of commercial purebred chicken lines for deleterious and functional variations, combining genotype and whole-genome sequence data. I provide a genomic perspective on deleterious and functional genetic variation in three egg-laying breeds, giving insight into the process of purifying selection, and the role of recombination for breeds under strong artificial selection. In addition, we report multiple putative functional coding variants in selective sweep regions, which are likely under positive selection.

In **chapter 5** I describe a unique example of allelic pleiotropy for a large deletion affecting two different genes in pigs. The deletion produces a truncated BBS9 protein, which subsequently results in enhanced growth rates in carrier animals. Intriguingly, a loss of function of this gene in human and mice leads to obesity. We show that fetal death in homozygotes, however, is not caused by an impaired *BBS9* gene, but by reduced expression of the downstream *BMPER* gene, an essential gene for normal foetal development. Finally, we provide strong evidence for balancing selection, given a higher selection index for carrier animals, resulting in an unexpected high frequency of this lethal allele in the population.

Chapter 6 describes loss of function mutations in essential genes that cause embryonic lethality in pigs, decreasing litter sizes by 15.1 to 21.6% in carrier-by-carrier matings. I first reflect on the effect of genetic drift on lethal recessive variants, showing that lethal alleles can reach allele frequencies up to 10% (20% carrier frequency) by genetic drift alone. Next, I describe in detail the loss-of-function mutations that impair essential genes leading to embryonic lethality in homozygous embryos. The causal mutations are of different type including two splice-site variants (affecting *POLR1B* and *TADA2A* genes), one frameshift (*URB1*), and one missense (*PNKP*) variant. Finally, I describe the impact of the lethals on population fitness, and its role in the heterosis effect observed for crossbred litters.

In **chapter 7** I describe a recessive 16-bp deletion in the *SPTBN4* gene causing severe myopathy and postnatal mortality in pigs. The deletion induces a frameshift and a premature stop codon, producing an impaired and truncated spectrin beta non-erythrocytic 4 protein (SPTBN4). The affected piglets are unable to walk and die within 24 hours after birth. This study shows how the growing resources of genomics data can aid in the identification of variants that result in piglet mortality, and lead to the subsequent prevention of carrier-by-carrier crosses. Finally, I aim to increase awareness among breeders and farmers of “hidden” genetic defects in the population, which helps to identify rare syndromes in breeding populations in the future.

Finally, in **chapter 8** I provide a framework to pinpoint likely causal variation and genes underlying important phenotypes in pigs. The variant prioritization method relies on the pig Combined Annotation Dependent Depletion (pCADD) scores, a machine learning method that provides impact scores to any possible substitution in the pig genome. I demonstrate the efficacy of the tool by reporting known and novel causal variants, of which many affect (non-coding) regulatory sequences associated with important phenotypes in pigs. Finally, the identified causal variants can be applied in breeding to improve genomic prediction.

Curriculum Vitae

About the author

Martijn Derks was born on the 13th of April 1991 in Wellerlooi, the Netherlands. He received his HAVO degree from the Valuascollege in Venlo. Martijn was always interested in computers, and biology was one of his favorite subjects at school. Hence, he chose to start a bachelor education Bioinformatics at the Hogeschool Arnhem Nijmegen in 2008. During his BSc internships he developed software for the analysis of different types of transcriptome data at The Centre for Molecular and Biomolecular Informatics (CMBI) and the department of Human Genetics in the Radboud University Medical Center in Nijmegen. After obtaining his bachelor degree in Nijmegen he continues to do a MSc Bioinformatics at the Wageningen University. During his MSc thesis he worked on the genome of the wintermoth, a model organism studied for its response to climate change, and subsequently did his internship at Human Nutrition working on the development of a software package to visualize different types of ~omics data. After he obtained his MSc degree in 2014, he started working as a research assistant in the Bioinformatics group at Wageningen University and the Animal Ecology group at the NIOO institute, in which he worked on various ~omics related projects in plants, animals, fungi, and bacteria.

In early 2016, Martijn started his PhD position in Animal Breeding and Genomics, to work on the "DelVar" Project within the STW-Breed4Food partnership consortium. He was attracted by the enormous amounts of data produced by our breeding partners and all the opportunities that comes with it. His research was initially focused on identifying lethal recessives in livestock populations, but shifted towards identifying other types of functionally important variation. In 2018 he visited the lab of Michel Georges at the GIGA institute in Liège. Martijn currently works as a lecturer at the University, and two days a week as junior researcher at Topigs Norsvin.

Peer reviewed publications

1. **Derks MFL**, Gjuvsland AB, Bosse M, Lopes MS, van Son M, Harlizius B, Tan BF, Hamland H, Grindflek E, Groenen MAM, et al., *Loss of function mutations in essential genes cause embryonic lethality in pigs.* Plos Genetics, 2019. 15(3).
2. **Derks MFL**, Harlizius B, Lopes MS, Greijdanus-van der Putten SWM, Dibbits B, Laport K, Megens HJ, and Groenen MAM, *Detection of a Frameshift Deletion in the SPTBN4 Gene Leads to Prevention of Severe Myopathy and Postnatal Mortality in Pigs.* Frontiers in Genetics, 2019. 10.
3. **Derks MFL**, Herrero-Medrano JM, Crooijmans RPMA, Vereijken A, Long JA, Megens HJ, and Groenen MAM, *Early and late feathering in turkey and chicken: same gene but different mutations.* Genetics Selection Evolution, 2018. 50.
4. **Derks MFL**, Lopes MS, Bosse M, Madsen O, Dibbits B, Harlizius B, Groenen MAM, and Megens HJ, *Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome.* Plos Genetics, 2018. 14(9).
5. **Derks MFL**, Megens HJ, Bosse M, Lopes MS, Harlizius B, and Groenen MAM, *A systematic survey to identify lethal recessive variation in highly managed pig populations.* Bmc Genomics, 2017. 18.
6. **Derks MFL**, Megens HJ, Bosse M, Visscher J, Peeters K, Bink MCAM, Vereijken A, Gross C, de Ridder D, Reinders MJT, et al., *A survey of functional genomic variation in domesticated chickens.* Genetics Selection Evolution, 2018. 50.
7. **Derks MFL**, Schachtschneider KM, Madsen O, Schijlen E, Verhoeven KJF, and van Oers K, *Gene and transposable element methylation in great tit (Parus major) brain and blood.* Bmc Genomics, 2016. 17.
8. **Derks MFL**, Smit S, Salis L, Schijlen E, Bossers A, Mateman C, Pijl AS, de Ridder D, Groenen MAM, Visser ME, et al., *The Genome of Winter Moth (Operophtera brumata) Provides a Genomic Perspective on Sexual Dimorphism and Phenology.* Genome Biology and Evolution, 2015. 7(8): p. 2321-2332.
9. Bortoluzzi C, Bosse M, **Derks MFL**, Crooijmans RPMA, Groenen MAM, and Megens HJ, *The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations.* Evolutionary Applications, 2020. 13(2): p. 330-341.
10. Bosse M, Megens HJ, **Derks MFL**, de Cara AMR, and Groenen MAM, *Deleterious alleles in the context of domestication, inbreeding, and selection.* Evolutionary Applications, 2019. 12(1): p. 6-17.

11. Cooper JW, Wilson MH, **Derks MFL**, Smit S, Kunert KJ, Cullis C, and Foyer CH, *Enhancing faba bean (Vicia faba L.) genome resources.* Journal of Experimental Botany, 2017. 68(8): p. 1941-1953.

12. Costa MCD, Artur MAS, Maia J, Jonkheer E, **Derks MFL**, Nijveen H, Williams B, Mundree SG, Jimenez-Gomez JM, Hesselink T, et al., *A footprint of desiccation tolerance in the genome of Xerophyiita viscosa.* Nature Plants, 2017. 3(4).

13. da Silva VH, Laine VN, Bosse M, Spurgin LG, **Derks MFL**, van Oers K, Dibbits B, Slate J, Crooijmans RPMA, Visser ME, et al., *The Genomic Complexity of a Large Inversion in Great Tits.* Genome Biology and Evolution, 2019. 11(7): p. 1870-1881.

14. de Carvalho JF, Oplaat C, Pappas N, **Derks MFL**, de Ridder D, and Verhoeven KJF, *Heritable gene expression differences between apomictic clone members in Taraxacum officinale: Insights into early stages of evolutionary divergence in asexual plants.* Bmc Genomics, 2016. 17.

15. Faddeeva-Vakhrusheva A, **Derks MFL**, Anvar SY, Agamennone V, Suring W, Smit S, van Straalen NM, and Roelofs D, *Gene Family Evolution Reflects Adaptation to Soil Environmental Stressors in the Genome of the Collembolan Orchesella cincta.* Genome Biology and Evolution, 2016. 8(7): p. 2106-2117.

16. Faddeeva-Vakhrusheva A, Kraaijeveld K, **Derks MFL**, Anvar SY, Agamennone V, Suring W, Kampfraath AA, Ellers J, Le Ngoc G, van Gestel CAM, et al., *Coping with living in the soil: the genome of the parthenogenetic springtail Folsomia candida.* Bmc Genomics, 2017. 18.

17. Gorter FA, **Derks MFL**, van den Heuvel J, Aarts MGM, Zwaan BJ, de Ridder D, and de Visser JAGM, *Genomics of Adaptation Depends on the Rate of Environmental Change in Experimental Yeast Populations.* Molecular Biology and Evolution, 2017. 34(10): p. 2613-2626.

18. Gross C, **Derks MFL**, Megens HJ, Bosse M, Groenen MAM, Reinders M, and de Ridder D, *pCADD: SNV prioritisation in Sus scrofa.* Genetics Selection Evolution, 2020. 52(1).

19. Marques A, Costa M-CD, Chathuri U, Jonkheer E, Zhao T, Schijlen E, **Derks MFL**, Nijveen H, Marcet-Houben M, Julca I, et al., *A blueprint of seed desiccation sensitivity in the genome of Castanospermum australe.* bioRxiv, 2019: p. 665661.

20. Pirovano W, Boetzer M, **Derks MFL**, and Smit S, *NCBI-compliant genome submissions: tips and tricks to save time and money.* Briefings in Bioinformatics, 2017. 18(2): p. 179-182.

21. Smit S, **Derks MFL**, Bervoets S, Fahal A, van Leeuwen W, van Belkum A, and van de Sande WWJ, *Genome Sequence of Madurella mycetomatis mm55, Isolated from a Human Mycetoma Case in Sudan.* Microbiology Resource Announcements, 2016. 4(3).

22. van Rijswijck IMH, **Derks MFL**, Abee T, de Ridder D, and Smid EJ, *Genome Sequences of Cyberlindnera fabianii 65, Pichia kudriavzevii 129, and Saccharomyces cerevisiae 131 Isolated from Fermented Masau Fruits in Zimbabwe.* Microbiology Resource Announcements, 2017. 5(14).

23. van Son M, Lopes MS, Martell HJ, **Derks MFL**, Gangsei LE, Kongsro J, Wass MN, Grindflek EH, and Harlizius B, *A QTL for Number of Teats Shows Breed Specific Effects on Number of Vertebrae in Pigs: Bridging the Gap Between Molecular and Quantitative Genetics.* Frontiers in Genetics, 2019. 10.

24. Wu Z, **Derks MFL**, Dibbits B, Megens HJ, Groenen MAM, and Crooijmans RPMA, *A Novel Loss-of-Function Variant in Transmembrane Protein 263 (TMEM263) of Autosomal Dwarfism in Chicken.* Frontiers in Genetics, 2018. 9.

25. Bortoluzzi C, Megens HJ, Bosse M, **Derks MFL**, Dibbits B, Laport K, Weigend F, Groenen MAM, and Crooijmans RPMA, *Parallel genetic origin of foot feathering in birds.* Molecular Biology and Evolution, 2020.

Training and Education

The Basic Package (3.0 credits)	
WIAS Introduction Day	2016
Course on philosophy of science and/or ethics	2019
Course on essential skills	2016

Disciplinary Competences (10.5 credits)	
Statistics for Life Sciences	2017
Pattern recognition	2017
PhD discussion group (Code Club)	2017-2019
Interactive post-graduate course on characterization, management and exploitation of genomic diversity in animals	2018
Statistics for Omics data analysis	2019

Professional Competences (9.0 credits)	
Project and Time Management	2016
Techniques for Writing and Presenting a Scientific Paper	2016
Reviewing a Scientific Paper	2016
Effective behaviour in your professional surroundings	2017
Writing grant proposals	2019
Presenting with Impact	2016
Orientation on teaching for PhD candidates	2017
PhD carousel	2016/2017
Supervising BSc and MSc thesis students	2018

Presentation Skills (4.0 credits)	
WIAS Science day (oral)	2017-2019
Plant and Animal Genome conference (oral)	2019-2020
BioSB conference(oral)	2017-2019
ESEB conference (oral)	2017
Zoology conference (oral)	2017
B-Wise symposium (oral)	2018

Teaching competences (6.0 credits)	
Genomics course (ABG-30306)	2016-2020
Supervising MSc student (4 times)	2016-2020
Review 2 proposals for Research Master Cluster	2016/2019
CodeClub (programming discussion group)	2017/2018
Lecture course next generation sequencing (LUMC)	2017
Lecture genomics course (ABG-30306)	2019

Total credits	**32.5**

Acknowledgements

During my four year PhD many people helped me in one way or another to complete this thesis. Only afterwards you start realizing how important these people are for your thesis and for your development as a person.

First I would like to thank my daily supervisor Hendrik-Jan. I admire how knowledgeable you are, and your passion for biology (and informatics) is unique. I really enjoyed our scientific and non-scientific discussions, especially also during the 'borrel' we had to discuss my propositions ;). Secondly I would like to thank my promotor Martien, you have been very supportive during my whole PhD and I am grateful that I can stay working in your group! I am sure we will have some very fruitful years to come! I also would like to thank Ole, Richard (I promise to improve my beerpong skills for the next PAG conference :P) and Mirte, you guys are the drivers of the 'genomics' group, and I very much appreciate all your input on my work.

In addition, I would like to thank all my ABG colleagues! I realized we are working in a wonderful group both scientifically and socially. I had great time during all our activities! Thanks for that! Special thanks to some of my fellow 'genomics' PhDs: Henri, Chiara, Jani, Vinicius, Maulik, Lim, Zhou, Langqing, Pascal, Maria (and others) for all nice scientific discussions, coffee chats, dinners, activities, and bovenal 'gezelligheid', I really enjoyed working with you guys! Thanks to Christian as well, we had a lot of interesting discussions together, but I think in the end our collaboration within the DelVar project has been very successful!! I would also like to thank Lisette, Fadma, Wilma and Rosilde for helping me with all kinds of things during my PhD, you are vital for the proper functioning of the ABG group.

During my PhD I spend some time in Belgium. I would specially like to thank Michel, Carole, and Ruth for helping me to have a very pleasant and fruitful stay in Liege, I really liked my time with you in Belgium!

I would also like to say a few words to people that were very important during the start of my academic carrier. Sandra Smit you have been vital for the start of my career, I really learned a lot from you during my MSc and the time after, and I enjoyed all the work we did together!! In addition, I would like to thank Dick de Ridder, Henk Hilhorst, Marcel Visser, and Kees van Oers for their confidence in me by having me in your group after my MSc, I really hope we see each other more often in the future.

I would also like to say a few words to the Thursday group (or Friday group nowadays :P), Heleen, Rico, Petra, Inge, Erik, Erik, Paul, Sevgin. I haven't been there a lot over the last few years but you were my closest friends in Wageningen and I always felt very welcome when I joined one of your activities. I will try to stick on for a few dinks now and then in the future. Moreover, I would like to thank Rico and Inge for being such nice housemates, we had a lot of fun together during our 'bieravondjes' in Bennekom!

I also want to acknowledge some people from the various breeding companies that were very interested in my work and helped me with all my 'breeding' related questions: Barbara, Marcos, Egbert, Maren, Arne, Eli, Marco, Katrijn, Jeroen, Addie, Erik, Randy. Without the help of our commercial breeding partners many of the science we do will not be possible. Special thanks to Barbara and Marcos!! Without the two of you I wouldn't have been able to achieve many of the things we did!!

I would also like to say a few words to my paranymphs. Dear Chiara and Jani, you two have become quite close friends with me over the last years, I really enjoyed sharing an office with you Chiara, and I admire your strong opinions about everything (although I not always agree :P). You both have one thing in common, you both find animals soooooo cute! You guys really love animals, and I believe that is something very important as an animal scientist. Jani, we have had lots of very nice coffee chats, and I really admire how open, social, and loving you are (I think I could learnt a bit from you in that sense). I wish you two all the best for the rest of your careers, I am confident that you will both manage that. Thank you for guiding me through this important moment of my life!

As letste woj ik gear mien familie bedanke vur al ollie steun. Pap en mam, gullie hebbe mej altied onvurwaardelik gesteunt ien mien keuzes, en ik vuul mej nog altied hielemol thuus bej ollie ien 'de Loi'. Ok mien zussen Evelien en Yvonne woj ik gear bedanke, we hebben t nie altied mekkelik gehad de latste joare, mar wej zien der as familie sterker uutgekomme, bedankt doarvur! Ok woj ik gear mien vrienden bedanke vur alle gezellige oavende ien de kroeg, mit carnavalswagen bouwen, zuupvakanties (haha), en ozze gezellige weekendjes weg. Lisanne, ik woj ow gear bedanke vur alle liefde, steun, en gezelligheid. Ik bin altied bleej um thuus te komme bej zun lieve meid, bedankt dat ge er altied vur meej ziet.

Colophon

This research was funded by the STW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality. This study was financially supported by NWO-TTW and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs-Norsvin. The use of the HPC cluster was made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR). The research visit of Martijn Derks to the Unit of Animal Genomics at the University of Liege (Belgium) was financially supported by WIAS.

The cover was designed by Lu Lu

Printed by DigiForce | Proefschriftmaken.nl