LUCAS LIMA VERARDO

GENE NETWORKS FROM GENOME WIDE ASSOCIATION STUDIES FOR PIG REPRODUCTIVE TRAITS

Thesis presented to the Animal Science Program of Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRAZIL
2015

LUCAS LIMA VERARDO

GENE NETWORKS FROM GENOME WIDE ASSOCIATION STUDIES FOR PIG REPRODUCTIVE TRAITS

Thesis presented to the Animal Science Program of Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Doctor Scientiae*.

APROVADA: 31 de julho de 2015.

Cláudio Lisias Mafra de Siqueira

Marco Antonio Machado

Ole Madsen

Fabyano Fonseca e Silva
(Coorientador)

Simone Eliza Facione Guimarães
(Orientadora)

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
V476g
2015

Verardo, Lucas Lima, 1983-
Gene networks from genome wide association studies for pig reproductive traits / Lucas Lima Verardo. – Viçosa, MG, 2015.
v, 170f. : il. (algumas color.) ; 29 cm.

Orientador: Simone Eliza Facioni Guimarães.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Suíno - Melhoramento genético. 2. Genoma. 3. Suíno - Reprodução. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-graduação em Zootecnia. II. Título.

CDD 22. ed. 636.40821

AKNOWLEDGEMENTS

Firstly, I would like to thank God for the care and comfort in every moment and for giving me strength to finish this study.

Sincere thanks also to:

The most important people in my life, my parents Márcio and Beth, also my grandma Neuza who always supported me and believed in me;

My fiancée Juliana for all love, patience, friendship and stay besides me all the time giving me support;

My brother and sister (Paulo and Liana) and all my family for giving me all support and friendship;

My friends from LABTEC (Animal Biotechnology Laboratory), Renata, Darlene, André, Carol, Margareth, Walmir, Mayara, Débora, Iuri, Karine, Leticia, Ivan, Ygor, Alessandra and Regina for continuous knowledge, learning exchange and pleasant workplace;

Federal University of Viçosa (UFV), in special to the Animal Science Department (DZO) and Genetics and Breeding group, for the opportunity of carrying out the course;

Wageningen University in special to the Animal Science Department, Animal Breeding and Genetics Group for the opportunity of carrying out my sandwich PhD, and all WUR colleagues;

Professor Simone Eliza Facioni Guimarães, my adviser, for her excellent supervision, competence, teachings, friendship and opportunities given to me during all these years since my Bachelors;

Professor Fabyano Fonseca for all his valuable teachings and patience on statistics lessons and his friendship;

Professor Paulo Sávio Lopes, for his valuable teachings, patience and friendship;

Professor Ole Madsen, my adviser at WUR, for his excellent work on given me directions and good discussions on my work;

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the financial support;

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PDSE) for the financial support during the sandwich PhD;

TABLE OF CONTENTS

RESUMO

VERARDO, Lucas Lima, D.Sc., Universidade Federal de Viçosa, Julho de 2015. **Redes gênicas a partir de estudos de associação genômica ampla para características reprodutivas de suínos.** Orientadora: Simone Eliza Facioni Guimarães. Coorientadores: Fabyano Fonseca e Silva e Paulo Sávio Lopes.

Características reprodutivas em suínos como numero de natimortos (SB), numero total de nascidos (TNB) e numero de tetos (NT) são amplamente incluídos em programas de melhoramento devido suas importâncias na indústria. Ao contrário da maioria dos estudos de associação, que consideram fenótipos contínuos com um enfoque Gaussiano, estas características são conhecidas como variáveis discretas, podendo assim, potencialmente seguir outras distribuições como a Poisson. Além disso, apesar de haver vários estudos de associação genômica ampla (GWAS) sendo realizados, somente alguns vem explorando os significados biológicos dos genes identificados nestes estudos. O presente trabalho, usando análises pós-GWAS, fornece uma valiosa fonte de informações sobre genes identificados a partir de estudos de associação para características reprodutivas. As analises de distribuição em modelos genômicos demonstraram a importância em considerar modelos de contagem para SB. Além do mais, diferentes grupos de SNPs e blocos de QTL relevantes entre e dentro de cada estudo foram identificados, direcionando para a possibilidade de diferentes grupos de genes estarem desempenhando funções biológicas relacionadas a uma única característica. Deste modo, destacamos que a diversidade genômica entre populações/ambientes deve ser observada em programas de melhoramento de modo que populações de referência especificas para cada população/ambiente sejam consideradas em estudos genômicos. Com base nestes resultados, nós demonstramos a importância das análises pós-GWAS aumentando o entendimento biológico de genes relevantes para características complexas.

ABSTRACT

VERARDO, Lucas Lima, D.Sc., Universidade Federal de Viçosa, July, 2015. **Gene networks from genome wide association studies for pigs reproductive traits.** Adviser: Simone Eliza Facioni Guimarães. Co-Advisers: Fabyano Fonseca e Silva and Paulo Sávio Lopes.

Reproductive traits in pigs, such as number of stillborn (SB), total number born (TNB) and number of teats (NT), are widely included in breeding programs due their importance to the industry. As opposite to most association studies that consider continuous phenotypes under Gaussian assumptions, these traits are characterized as discrete variable, which could potentially follow other distributions, such as the Poisson. In addition, even though many genome wide association studies (GWAS) have been performed, only a few studies have explored biological meanings of genes identified. The present study provided a rich information resource about genes identified using genome wide association approaches for reproductive traits. The distribution analyses in genomic models, highlighted the importance in consider counting models for SB. Moreover, different sets of relevant SNPs and QTL blocks across and within the studies were identified leading to the possibility of different set of genes playing biological roles related to a single complex trait. Thereby, we highlighted the genomic diversity across population/environments to be observed in breeding programs in such a way that population/environments specific reference populations might be considered in genomic analyses. Based on these results, we demonstrated the importance of post-GWAS analyses increasing the biological understanding of relevant genes for complex traits.

Chapter 1

GENERAL INTRODUCTION

Reproductive traits

Reproductive traits, such as number of stillborn (SB), total number born (TNB) and number of teats (NT), are widely included in the selection indices of pig breeding programs due their importance to the pig industry. In pig production systems, around 30% of culling has primarily been because of reproductive problems (Stalder et al. 2004). The SB is directly affected by the TNB trait (Blasco et al., 1995) and temporal gene effects in different parities (Onteru et al., 2012). In a biological level, it has been shown that kidney diseases and diabetes in the mother and congenital heart disease in the fetuses are some of the main cause of stillbirth occurrence in humans (Nevis et al., 2011; Mathiensen et al., 2011 and Lee et al., 2001). In pigs, another factor acting on TNB phenotypic variance is the environment (Knap and Su, 2008). Nowadays, assisted reproductive techniques, such as artificial insemination, have allowed the sire distribution across multiple environments highlighting the genotype x environment interaction (GxE) relevance in pig production. Thus, a better biological understanding of these traits is still needed to improve selection against SB or for an increased TNB in pig breeding.

Number of teats (NT) is a trait with a large influence on the mothering ability of sows (Hirooka et al., 2001), being a limiting factor for the increased number of weaned piglets. Biologically, the development of embryonic mammary glands require the coordination of many signaling pathways to direct the cell shape changes, cell movements, and cell–cell interactions necessary for proper morphogenesis of mammary glands (Hens and Wysolmerski, 2005). In addition, number of vertebrae, that will determine the body length of the sow, may also have a direct relation with the final NT

observed in pigs (Ren et al., 2012). As these traits are directly involved with higher production and welfare of piglets, several association studies have been performed for SB, TNB and NT (Onteru et al., 2012; Uimari et al., 2011 and Schneider et al., 2012).

Genome Wide Association Studies - GWAS

In GWAS, these traits have been presumed normally distributed, which may not be true. All three, SB, TNB and NT, are measured as count variables, and therefore they may follow discrete distributions such as Poisson. Although Poisson distribution has already been implemented in animal breeding in the context of traditional mixed models (Perez-Enciso et al., 1993 and Ayres et al., 2013) and Quantitative Trait Loci (QTL) mapping (Cui et al., 2006 and Silva et al., 2011), there are no reports of GWAS using this approach. Poisson models can be performed using a Bayesian Markov chain Monte Carlo (MCMC) approach (Hadfield JD and Nakagawa, 2010). Applying Poisson models to GWAS, using a traditional mixed model, is possible to assume all markers simultaneously by using the genomic relationship matrix (Van Raden, 2008), as preconized in genomic best linear unbiased prediction (GBLUP). GBLUP has been widely used in genome wide selection (GWS) on which the GEBV vector, from each MCMC iteration, can be directly converted into SNP marker effects vector (Stranden and Garrick, 2009 and Wang et al., 2012). Based on this, samples of posterior distribution for each particular SNP effect are generated at each cycle and significance tests, based on Highest Posterior Density (HPD) intervals, can be performed in order to identify the most relevant markers. Studies have also adopted the HPD intervals to infer directly on marker effects significance (Li et al., 2012 and Ramírez et al., 2014). Additionally, the posterior probability (PPN0) of the estimated effect being lower than 0 (for negative effects) or greater than 0 (for positive effects) as proposed by Ramírez et

al. (2014) and Cecchinato et al. (2014) can be also used to report marker effect significance under a Bayesian approach.

Besides the distribution effect on GWAS, populations and environment differences have been reported (Veroneze et al., 2014 and Silva et al., 2014). Nowadays, assisted reproductive techniques, such as artificial insemination, have allowed the sire distribution across multiple environments highlighting the genotype x environment interaction (GxE) relevance in pig production. GxE studies, using genomic information, are important in view of the promising future of genomic wide selection (GWS) aiming to increase the genetic gain in pigs. These analyses can be accessed using different methodologies, such as multitrait models and its generalizations (Meyer, 2009), and also random regression reaction norm models (Kolmodin et al., 2002, Calus and Veerkamp, 2003; Cardoso and Tempelman, 2012). Under a random regression reaction norm models approach, it is possible to estimate the additive genomic breeding values (GEBV) for animals over different environmental classes, and furthermore it is also possible to estimate single nucleotides polymorphism (SNP) effects over such classes (Silva et al., 2014).

These approaches, allow us to perform a genome wide association study (GWAS) defining relevant SNP in different populations or environment class. Thus, a genetic dissection of complex phenotypes, through candidate genes network derived from relevant SNP might provide a better biological understanding of these traits.

Post-GWAS (Gene networks)

Currently, a point that deserves to be highlighted in GWAS is the genetic dissection of complex phenotypes through candidate genes network derived from SNP. There are relevant studies involving these networks in human disease (Liu et. al., 2011) and

puberty related traits in cattle (Fortes et al., 2011; Reverter and Fortes, 2013). However, in the pig this approach has not been exploited yet. In summary, these networks can be performed using the genes related to significant SNPs, to examine the sharing of pathways and functions involving these genes. In addition, transcription factors (TF) analyses can be implemented.

It has been shown that TF have relations with important traits in pigs, besides their roles in regulate the gene expression. Yu et al. (1995) demonstrated that the TF PIT1 might be related with carcass traits. Chen et al. (2008) showed the SREBF1 relation with regulation of muscle fat deposition; and Markljung et al. (2009) cited that the TF ZBED6 is related with muscle development. Providing evidence for the interaction between TF and its predicted targets with a further gene-TF network construction must be helpful in point out the most probable group of candidate genes. Studies involving gene networks in puberty related traits for cattle have been performed (Fortes et al., 2011; Reverter and Fortes, 2013); in pigs, this approach has began to be more exploited with this work.

Objectives and outliers

This thesis describes studies conducted on GWAS and Post-GWAS applied to reproductive traits in different scenarios. At the second chapter, we aimed to present a full Bayesian treatment of SNP association analysis for number of teats (NT) assuming Gaussian and Poisson distributions for this trait. Under this framework, significant SNP effects were identified by hypothesis tests using 95% highest posterior density intervals. Moreover, we used these SNPs to construct an associated candidate genes network, and TF analyses, aiming to explain one possible genetic mechanism behind the referred trait. At the third chapter, we also performed and tested a Bayesian treatment of a

GWAS model assuming Poisson and Gaussian distribution. But, for this study, we used SB and NT data in a large commercial pig line. Moreover, we also used the significant SNPs to obtain the related genes and generate gene-transcription factor networks aiming at exploring biological roles behind the considered traits and pointing out the most probable candidate genes.

At the fourth chapter, we present a gene network analysis across two dam lines based on a Bayesian GWAS for NT. For each line separately, genes linked to significant SNPs were identified. These genes were used to construct a combined network and also to detect related TF in each line, aiming to obtain the most likely candidate genes for NT in each line, followed by a comparative analysis to assess (dis)similarities at marker and gene level across the lines. On the fifth chapter, we present a gene network analysis for total number born trait (TNB) in pigs across different environment levels using results from a random regression reaction norm models GWAS approach. Under this framework, relevant SNP were identified as their linked genes for each environment level. These genes were used to construct a combined network and also search for related TF across groups of levels, aiming to obtain a set of candidate genes in each group based on their shared TF binding sites. Therefore, we were able to observe the similarities (or not) at marker and gene level across different environment groups.

REFERENCES

Ayres DR, Pereira RJ, Boligon AA, Silva FF, Schenkel FS, Roso VM, et al.: **Linear and Poisson models for genetic evaluation of tick resistance in cross-bred Hereford x Nellore cattle**. *J Anim Breed Genet* 2013, **130**(6):417-424.

Blasco A, Bidanel JP, HaleyCS: **Genetics and neonatal survival**. In: Varley MA, editors. *The neonatal pig: development and survival*. CAB Int, Wallingford, UK; 1995. p. 17-38.

Calus MPL, Veerkamp R: **Estimation of environmen- tal sensitivity of genetic merit for milk production traits using a random regression model**. *J Dairy Sci* 2003, **86**:3756–3764.

Cardoso FF, Tempelman RJ: **Linear reaction norm models for genetic merit prediction of Angus cattle under geno- type by environment interaction**. *J Anim Sci* 2012, **90**(7):2130–2141.

Cecchinato A, Ribeca C, Chessa S, Cipolat-Gotet C, Maretto F, Casellas J, Bittante G: **Candidate gene association analysis for milk yield, composition, urea nitrogen and somatic cell scores in Brown Swiss cows.** *Animal* 2014, **7**:1-9.

Chen J, Yang XJ, Xia D, Wegner J, Jiang Z, Zhao RQ: **Sterol regulatory element binding transcription factor 1 expression and genetic polymorphism significantly affect intramuscular fat deposition in the longissimus muscle of Erhualian and Sutai pigs**. *J Anim Scie* 2008, **86**(1):57-63.

Cui Y, Kim D-Y, Zhu J: **On the Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci With Count Data**. *Genetics* 2006, **174**(4):2159-2172.

Fortes MR, Reverter-Gomez T, Hiriyur-Nagaraj S, Zhang Y, Jonsson N, Barris W, et al.:**A SNP-derived regulatory gene network underlying puberty in two tropical breeds of beef cattle**. *J Anim Sci* 2011, **89**:1669-1683.

Hadfield JD, Nakagawa S: **General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters.** *J Evol Biol* 2010, **23**:494-508.

Hens JR, Wysolmerski JJ: **Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland**. *Breast Cancer Res* 2005, **7**(5):220.

Hirooka H, de Koning DJ, Harlizius B, van Arendonk JA, Rattink AP, Groenen MA, et al.: **A whole-genome scan for quantitative trait loci affecting teat number in pigs**. *J Anim Sci* 2001, **79**(9):2320-2326.

Knap PW, Su G: **Genotype by environment interaction for litter size in pigs as quantified by reaction norms analysis**. *Animal* 2008, **2**:1742–1747.

Kolmodin R, Strandberg E, Madsen P, Jorjani H: **Genotype by environment interaction in Nordic dairy cattle studied using reaction norms**. *Acta Agric Scand* 2002, **52**:11–24.

Lee K-sun, Khoshnood B, Chen L, Stephen NW, Cromie JW, Mittendorf RL: **Infant mortality from congenital malformations in the United States, 1970 –1997**.*Obstet Gynecol* 2001, **98**:620-627.

Li Z, Gopal V, Li X, Davis JM, Casella G: **Simultaneous SNP identification in association studies with missing data**. *Ann App Stat* 2012, **6**(2):432-456.

Mathiesen ER, Ringholm L, Damm P: **Stillbirth in diabetic pregnancies**. *Best Pract Res Clin Obstet Gynaecol* 2011, **25**(1):105-111.

Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, et al.: **ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth**. *PLoS biology* 2009, **7**(12):2738.

Nevis IF, Reitsma A, Dominic A, McDonald S, Thabane L, Akl EA, et al.: **Pregnancy Outcomes in Women with Chronic Kidney Disease: A Systematic Review**. *Clin J Am Soc Nephrol* 2011, **6**(11):2587-2598.

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

Perez-Enciso M, Tempelman RJ, Gianola D: **A comparison between linear and Poisson mixed models for litter size in Iberian Pigs**. *Livest Prod Sci* 1993, **35**:303.

Ramírez O, Quintanilla R, Varona L, Gallardo D, Díaz I, Pena RN, Amills M: **DECR1 and ME1 genotypes are associated with lipid composition traits in Duroc pigs**. *J Anim Breed Genet* 2014, **131**(1):46-52.

Ren DR, Ren J, Ruan GF, Guo YM, Wu LH, Yang GC, et al.: **Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc × Chinese Erhualian intercross resource population**. *Anim Genet* 2012, **43**(5):545-551.

Reverter A, Fortes MRS: **Building single nucleotide polymorphism-derived gene regulatory networks: Towards functional genomewide association studies**. *J Anim Sci* 2013, **91**:530-536.

Schneider JF, Rempel LA, Rohrer GA: **Genome-wide association study of swine farrowing traits. Part I: Genetic and genomic parameter estimates**. *J Anim Sci* 2012, **90**:3353-3359.

Silva FF, Mulder HA, Knol EF, Lopes MS, Guimarães SEF, Lopes PS et al.: **Sire evaluation for total number born in pigs using a genomic reaction norms approach**. *J Anim Sci* 2014, **92**(9), 3825-3834.

Silva KM, Knol EF, Merks JWM, Guimarães SEF, Bastiaansen JWM, van Arendonk JAM, et al.: **Meta-analysis of results from quantitative trait loci mapping studies on pig chromosome 4**. *Anim Genet* 2011, **42**:280-292.

Stranden I, Garrick DJ: **Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit**. *J Dairy Sci* 2009, **92**:2971–2975.

Uimari P, Sironen A, Sevón-Aimonen ML: **Whole-genome SNP association analysis of reproduction traits in the Finnish Landrace pig breed**. *Genet Sel Evol* 2011, **43**:42.

Van Raden PM: **Efficient methods to compute genomic predictions**. *J Dairy Sci* 2008, **91**(11):4414-4423.

Veroneze R, Bastiaansen J, Knol EF, Guimarães S, Silva FF, Harlizius B, et al.: **Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations**. BMC Genet 2014, **15**(1):126.

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

Yu TP, Tuggle CK, Schmitz CB, Rothschild MF: **Association of PIT1 polymorphisms with growth and carcass traits in pigs**. *J Anim Sci* 1995, **73**(5):1282-1288.

Chapter 2

Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs

L.L. Verardo[1], F.F. Silva[1], L.Varona[2], M.D.V. Resende[3], J.W.M. Bastiaansen[4], P.S. Lopes[1] and S.E.F. Guimarães[1]

[1] Department of Animal Science, Universidade Federal de Viçosa - UFV, Viçosa - MG, Brazil
[2]Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, Zaragoza, Spain
[3]Embrapa Florestas, Colombo - PR, Brazil.
[4]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands,

Journal of Applied Genetics (2015) 56:123-132

Abstract

The genetic improvement of reproductive traits such as the number of teats is essential to the success of pig industry. As opposite to most SNP association studies that consider continuous phenotypes under Gaussian assumptions, this trait is characterized as discrete variable, which could potentially follow other distributions, such as the Poisson. Therefore, in order to access the complexity of a counting random regression considering all SNPs simultaneously as covariate under a GWAS modeling, the Bayesian inference tools become necessary. Currently, another point that deserves to be highlighted in GWAS is the genetic dissection of complex phenotypes through candidate genes network derived from significant SNPs. We present a full Bayesian treatment of SNP association analysis for number of teats assuming alternatively Gaussian and Poisson distributions for this trait. Under this framework, significant SNP effects were identified by hypothesis tests using 95% highest posterior density intervals. These SNPs were used to construct associated candidate genes network aiming to explain the genetic mechanism behind this reproductive trait. The Bayesian model comparisons based on deviance posterior distribution indicated the superiority of Gaussian model. In general, our results suggest the presence of 19 significant SNPs, which mapped 13 genes. Besides, we predicted gene interactions through networks that are consistent with the mammals known breast biology (e.g., development of prolactin receptor signaling, and cell proliferation), captured known regulation binding sites, and provided candidate genes for that trait (e.g., TINAGL1 and ICK).

Keywords: Reproductive traits, counting data, SNP association, genes.

Introduction

An important trait related to the success of pig reproduction is the number of teats. It reflects directly the mothering ability of sows (Hirooka et. al., 2001), which is a limiting factor for the increased number of weaned piglets. This trait is known to have a low to medium heritability (Clayton et al., 1981; and McKay and Rahnefeld, 1990), thus the use of genome-wide association studies (GWAS) can be useful to search for chromosomal regions that can help to explain the genetic architecture of this complex trait.

At present, many studies has been done using GWAS for reproductive traits in pigs (Uimari, et. al., 2011; Onteru et. al., 2011 and Schneider et. al., 2012). However, these studies have not pointed out to the discrete nature of these traits, which are usually considered as counting variables (i.e., number of teats, number of stillborn and number of weaned, among others). This kind of traits could potentially follow an appropriate discrete distribution, such as Poisson. Although this has already been implemented in animal breeding in the context of mixed models (Perez-Enciso et al., 1993; Ayres et al., 2013; Varona and Sorensen, 2010) and Quantitative Trait Locus (QTL) detection (Cui et. al., 2006; Silva et. al., 2011), there are no reports of GWAS under a Poisson distribution approach. Therefore, to solve problems related to the complexity of a Poisson random regression model in a GWAS context, the Bayesian inference becomes necessary.

Currently, another point that deserves to be highlighted in GWAS is the genetic dissection of complex phenotypes through candidate genes network derived from significant single nucleotide polymorphism (SNP) for different traits. There are relevant studies involving these networks in human disease (Liu et. al., 2011) and puberty related traits in cattle (Fortes et al., 2011; Reverter and Fortes, 2013).

However, in the pig this approach has not been exploited yet. In summary, these networks can be performed using the genes symbols related to significant SNPs, and can be used to examine the process of shared pathways and functions involving these genes. Besides, an *in silico* validation for these studies trough Transcription Factors (TF) analyses can be performed.

Toward this orientation, we aimed to present a full Bayesian treatment of SNP association analysis for number of teats assuming Gaussian and Poisson distributions for this trait. Under this framework, significant SNP effects were identified by hypothesis tests using 95% highest posterior density intervals. Moreover, we used these SNPs to construct an associated candidate genes network, and TF analyses, aiming to explain one possible genetic mechanism behind the referred trait.

Material and Methods

Experimental population and phenotypic data

The phenotypic data was obtained from the Pig Breeding Farm of the Department of Animal Science, Universidade Federal de Viçosa (UFV), MG, Brazil. A three-generation resource population was created and managed as described by Band et al. (2005a). Briefly, two local breed Piau grandsires were crossed with 18 granddams from a commercial line composed of Large White, Landrace and Pietrain breeds, to produce the F1 generation from which 11 F1 sires and 54 F1 dams were selected. These F1 individuals were crossed to produce the F2 population, of which 345 animals were phenotyped for number of teats.

DNA extraction, genotyping and SNP selection

DNA was extracted at the Animal Biotechnology Lab from Animal Science Department of Universidade Federal de Viçosa. Genomic DNA was extracted from white cells of parental, F1 and F2 animals, more details can be found in Band et, al.,

2005b.The low-density(Habier et al., 2009) customized SNPChipwith 384 markers was based on the Illumina Porcine SNP60 BeadChip (San Diego, CA, USA, Ramos et. al., 2009). These SNPs were selected according to QTL positions previously identified on this population using meta-analyses (Silva et al., 2011) and fine mapping (Hidalgo et al., 2013). From these, 66 SNPs were discarded for no amplification, and from the remaining 318 SNPs, 81 were discarded due to a minor allele frequency (MAF) < 0.05. Thus, 237 SNPs markers were distributed as follows: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (30), SSC17 (25) and SSCX (13), being the average distance within each chromosome, respectively: 5.17, 2.37, 2.25, 3.93, 2.68 and 11.0 Mb.

Statistical modeling and computational features

It was proposed a hierarchical Bayesian multiple regression model considering two different distributions for the data, Gaussian and Poisson. In these models, all SNPs were fitted simultaneously, analogously to Bayesian models used in genome wide selection (Meuwissen et al., 2001). However, given the small number of markers in the present study, we considered an improvement by inclusion of covariance between SNP effects as unknown parameters. In the first case, when the Gaussian distribution was assumed for the phenotypes, the following regression model was considered:

$$y_i = \mu + sex + batch + hal + \sum_{k=1}^{237} x_{ik}\beta_k + e_i, \qquad (1)$$

Where y_i is the phenotypic observation of animal i (i = 1, 2,…, 345); μ is the general mean; sex, batch and halothane (hal) gene genotype are the fixed effects; β_k is the allelic substitution effect of marker k and e_i is the residual term $e_i \sim N(0, \sigma_e^2)$.In this model, the covariate x_{ik} takes the values 2, 1and 0, respectively to the SNP genotypes AA, Aa and aa at each locus k. It was assumed a multivariate normal distribution for

the SNP effects vector, $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_{237}]'$, $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ the covariance matrix between markers. Since this matrix is considered as an unknown parameter, its prior was given by an inverted Wishart distribution, $\boldsymbol{\Sigma} \sim IW(\nu, \boldsymbol{\Sigma}_0)$, in which ν is the shape parameter and $\boldsymbol{\Sigma}_0$ the scale matrix. To incorporate a prior knowledge about the true covariance between SNP effects, we proposed to use the linkage disequilibrium (LD) matrix as $\boldsymbol{\Sigma}_0$ matrix. Thus, this matrix contained r^2 values provided by *snpgdsLDMat* function of package *SNPRelate* (Zheng et al., 2012) of R software (R Core Team, 2013). For the fixed effects and residual variance, non-informative (Uniform) and inverse Gamma, $\sigma_e^2 \sim IGamma(a, 1/b)$ distributions were assumed, respectively.

The second approach assumed a Poisson distribution, $y_i \sim Po(\lambda_i)$, and the model (1) was rewritten under a generalized linear model (2) approach, in which λ_i is the Poisson mean and $\log(\lambda_i)$ a latent variable defined from the canonical (logarithm) link function as follow:

$$\log(\lambda_i) = \mu + sex + batch + hal + \sum_{k=1}^{237} x_{ik}\beta_k + e_i. \quad (2)$$

The prior distributions assumed for the parameters of model (2) were the same as in model (1). However, the latent variables are now considered also as unknown parameters, do not having a recognizable conditional distribution. Thus, the Metropolis–Hastings algorithm was required for the implementation of the MCMC algorithm.

The models (1) and (2) were implemented, respectively, in the functions *MCMChregress* (MCMC for the hierarchical Gaussian linear regression model) and *MCMChpoisson* (MCMC for the hierarchical Poisson linear regression model) of *MCMCpack* (Martin et. al., 2011) R package. It was considered a total of 100,000 iterations with a burn-in period and sampling interval (thin) of 50,000 and 5 iterations,

respectively. All used codes are available in supplementary material (ESM_1.pdf), and the real data set can be requested directly with the authors. The convergence of MCMC chains was verified by Geweke test using *boa* (Bayesian output analysis) R package (Smith, 2007).

Models were compared by using the posterior distribution of the deviance $P(D_M)$ provided by a particular M model. For the Gaussian model, each value of this distribution is obtained directly by $D_G^{(j)} = -2 \log(\prod_{i=1}^{345} P(y_i|\boldsymbol{\theta}^{(j)}))$, in which $\prod_{i=1}^{345} P(y_i|\boldsymbol{\theta}^{(j)})$ is the value for the likelihood function considering the set of parameter estimates ($\boldsymbol{\theta}^{(j)}$) at each MCMC iteration j. Similarly, for the Poisson model, the values came from $D_P^{(j)} = -2 \log(\prod_{i=1}^{345} P(y_i|\lambda_i^{(j)}))$, being $\lambda_i^{(j)}$ the estimate of Poisson mean, i.e. the exponential of the latent variable $\log(\lambda_i)$ generated by Metropolis-Hastings algorithm. Thus, the random draws from posterior distributions of the deviance for both models, $P(D_G)$ and $P(D_p)$, were used to simulate the distribution of deviance difference (Lorenzo-Bermejo et al. 2011) given by $P(D_{G-P})$. Once obtained this distribution, it was possible to propose a hypothesis test based on HPD (Highest Posterior Density) interval for the deviance difference. In this context, knowing that lower deviance values indicate better fitting model, if the interval contains only negative values, the Gaussian model is indicated as the best one. On the other hand, an interval containing only positive values implies the best fit of Poisson model.

Although the used SNPs are at pre-identified QTL positions in this population as previously cited, therefore explaining a large amount of total additive genetic variance, the polygenic effect $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, also was included in models (1) and (2) to point out for some eventual portion of variance that was not captured by markers.

In order to add this effect, we used the methodology proposed by Harville and Callanan (1989) and Vazquez et al. (2010), which is a reparametrization of the genetic values vector (**u**) by assuming the traditional relationship matrix (**A**) as a diagonal matrix. This strategy is useful when using computational tools in which it is impossible to specify the **A** matrix directly as the covariance of random effects, as is the case of *MCMCpack* used in the present study. In summary, this methodology is based in the Cholesky decomposition of **A** matrix (**A** = **LL′)** whose factor **L** is used to reparametrize the incidence matrix of random effects (**Z**), $\mathbf{Z}^{*}$=**ZL**, implying in $\mathbf{u}^{*} \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2)$, being **u*** a reparametrized vector of genetic breeding values. Under this approach, it is possible the addition of the individual random effect directly in the models (1) and (2), whose solution ($\hat{\mathbf{u}}^{*}$) must be used to obtain the original vector of breeding values, which is given by $\hat{\mathbf{u}} = \mathbf{L}\hat{\mathbf{u}}^{*}$**.** The deviance posterior distributions were used to compare the models with and without the polygenic effect in order to verify its real influence in the studied phenotype.

Once identified the best model, the significance of SNP effects also can be obtained directly through 95% HPD intervals. Under this approach, if the interval contains the value zero, the SNP effect is non-significant. These intervals were constructed for each marker, so that the chromosome positions of the significant SNPs were used for identifying genes influencing the analyzed trait.

Genes Network and Regulatory Sequence Analysis

Initially, we identified the SNP related genes (the genes which had a SNP in it sequence or up to 2500 bp before the gene start or after the end) at dbSNPNCBI web site (http://www.ncbi.nlm.nih.gov/SNP/) through significant SNPs location and related gene symbol. For genes that did not have a pig symbol, we used the human related identifier. The GeneCards web site (http://www.genecards.org/) and the

program TOPPCLUSTER (http://toppcluster.cchmc.org/) were used to obtain the genes relationship as there functional GeneOntology (GO). Thus, it was possible to identify the biological mechanisms, pathways and functions involving them. The application Cytoscape (www.cytoscape.org/) was used to visualize and edit the identified network.

Providing evidence for the interaction between the TF and its predicted targets via regulatory sequence analysis serves as an *in silico* validation for the TF–target interactions in the SNP genes network (Fortes et. al., 2010). Here we used the TFM-Explorer (http://bioinfo.lifl.fr/TFM/TFME/), a freely available program. This program takes a set of gene sequences, and searches for locally overrepresented transcription factor binding sites (TFBS) using weight matrices from JASPAR database to detect all potential TFBS, and extracts significant clusters (region of the input sequences associated with a factor) by calculating a score function. This score threshold is chosen to give a *P*-value equal or better to 10^{-3} for each position for each sequence such as described in Touzet and Varré (2007). The top TF related (P-value<0.001) were identified and for the three most represented (according to P-value) we construct a network with their interactions (TF-target) and the gene ontology using the application Cytoscape and collecting information at GeneCards® website.

Results

Statistical Analyses

The results of model comparison between both Gaussian and Poisson approach identify the Gaussian model as the one with best fit, since presented a lower deviance values than Poisson model (Fig. 1). Considering the polygenic effect in Bayesian GWAS models, the analysis of the deviance posterior distribution indicated

that there is no gain in the models fitting quality when including this effect as shown on supplementary figure ESM_2.pdf. Using the results from the best model (Gaussian distribution), we could than identify 19 significant SNPs using a hypothesis test based on 95% HPD interval for their effects (Table 1).

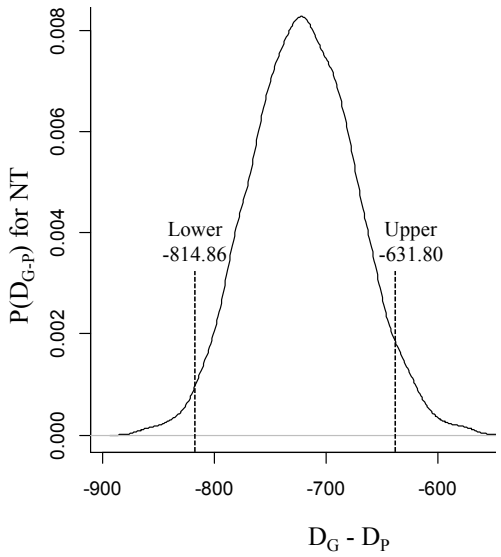


Fig. 1 Distribution of deviance difference, P(DG-P), between Gaussian and Poisson GWAS models fitted to number of teats (NT) in pigs. DG and DP are the estimated Gaussian and Poisson models deviance, respectively.

Table 1 Significant SNPs for the trait number of teats in pigs, their positions in base pairs (bp) at pig chromosome (chr) with their related genes, posterior mean and 95% HPD (Highest Posterior Density) intervals for SNP effects.

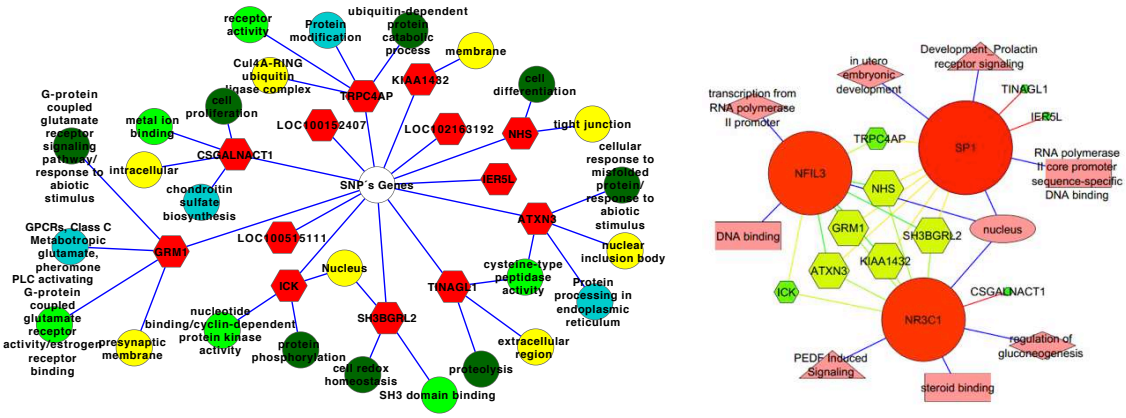
					95% HPD interval	
SNP	**Position (bp)**	**chr**	**Genes[a]**	**Posterior mean**	**Lower**	**Upper**
ALGA0001557	21576278	1	GRM1	-0.4551805	-0.7785922	-0.12447637
ALGA0004074	75728908	1	LOC100515111	-0.5026433	-0.9015982	-0.08779558
ALGA0004774	97143269	1	SH3BGRL2	-0.6260198	-1.0934972	-0.14063153
ALGA0007908	242165688	1	KIAA1432	0.4936047	0.09634696	0.9179789
ALGA0010677	303486951	1	IER5L	-0.5267897	-1.0204636	-0.02172639
ALGA0024439	34713731	4	-	0.5115040	0.09656826	0.9618503
ALGA0026100	84090680	4	-	0.6282089	0.06594756	1.1671385
ALGA0027472	100262783	4	-	0.3934344	0.01607044	0.8147838
ALGA0027647	114662647	4	-	0.4305747	0.01093520	0.8081567
ALGA0028649	129467665	4	-	-0.4954020	-0.9146217	-0.09985800
ALGA0039880	30978428	7	TINAGL1	-0.4930740	-0.8402710	-0.17129009
ALGA0043403	92021173	7	LOC100152407	-0.6302710	-1.1360743	-0.13179192
ALGA0044983	120223503	7	ATXN3	0.6438884	0.01955658	1.2668060
ALGA0045990	134422073	7	ICK	-0.4388611	-0.8196542	-0.08716361
ALGA0048133	75648737	8	-	0.5372867	0.19206533	0.8861433
ALGA0093254	10087010	17	CSGALNACT1	0.3899252	0.01529562	0.7985949
ALGA0094092	31195099	17	LOC102163192	0.3436231	0.05611619	0.6329979
ALGA0094911	43584674	17	TRPC4AP	0.5159049	0.14614182	0.8930050
ASGA0080951	15127052	x	NHS	-0.4990405	-0.8496433	-0.13886008

[a] GRM1: glutamate receptor, metabotropic 1; SH3BGRL2: SH3 domain-binding glutamic acid-rich-like protein 2-like; IER5L: immediate early response gene 5-like protein-like; TINAGL1: tubulointerstitial nephritis antigen-like; ICK: intestinal cell (MAK-like) kinase; CSGALNACT1: Chondroitin Sulfate N-Acetylgalactosaminyltransferase 1; KIAA1432: KIAA1432; ATXN3: ataxin 3; TRPC4AP : transient receptor potential cation channel, subfamily C, member 4 associated and NHS: Nance-Horan Syndrome.

Genes Network and Regulatory Sequence Analysis

Besides, from significant SNPs identified we could find 13 genes which have those polymorphisms in their sequences or close, they are: GRM1, LOC100515111, LOC100510992 (SH3BGRL2), LOC100620589 (IER5L), LOC100514061 (TINAGL1), ICK, KIAA1432, ATXN3, LOC100152407, LOC102166124 (CSGALNACT1), LOC102163192, NHS and TRPC4AP. To understand the functions of these genes, we collected information about their biological process, cellular component and molecular function in the Gene Ontology (GO). Furthermore, using the application TOPPCLUSTER, we were able to identify metabolic pathways and interaction based on human gene names as described in Table 2. With all genes

founded we construct a network with their pathway, biological process, molecular function and cellular component (Fig. 2a).



a b

Fig. 2 Functional networks and their interactions for number of teats trait. (a) The relationships between 13 genes (in octagon red) and their related subnets: pathway (in blue), biological process (in dark green), molecular function (in light green), and cellular component (in yellow). (b) Transcription Factor (TF) network showing three TF: NR3C1, NFIL3 and SP1(circles shape) with *in silico* validated targets (hexagon nodes), their node color scale corresponds to network analyses (cytoscape) score where red nodes represent higher edges density. Pink nodes are the TF related pathways (triangle), biological process (diamond), molecular function (rectangle), and cellular component (elliptic).

Table 2 Genes related to number of teats in pigs represented in the network explaining their pathway, biological process, molecular function and cellular component.

Genes	Hs Symbol[a]	Gene Description	Pathway	GO: Biological Process	GO: Molecular Function	GO: Cellular Component
GRM1	GRM1	glutamate receptor, metabotropic 1	GPCRs, Class C Metabotropic glutamate, pheromone	G-protein coupled glutamate receptor signaling pathway/ response to abiotic stimulus	PLC activating G-protein coupled glutamate receptor activity/estrogen receptor binding	presynaptic membrane
LOC100515111	-	uncharacterized	-	-	-	-
LOC100510992	SH3BGRL2	SH3 domain-binding glutamic acid-rich-like protein 2-like	-	cell redox homeostasis	SH3 domain binding	Nucleus
LOC100620589	IER5L	immediate early response gene 5-like protein-like	-	-	-	-
LOC100514061	TINAGL1	tubulointerstitial nephritis antigen-like	-	proteolysis	cysteine-type peptidase activity	extracellular region
LOC100152407	-	uncharacterized	-	-	-	-
ICK	ICK	intestinal cell (MAK-like) kinase	-	protein phosphorylation	nucleotide binding/cyclin-dependent protein kinase activity	Nucleus
LOC102166124	CSGALNACT1	Chondroitin Sulfate N-Acetylgalactosaminyltransferase 1	chondroitin sulfate biosynthesis	cell proliferation	metal ion binding	intracellular
LOC102163192	LOC102163192	-	-	-	-	-
KIAA1432	KIAA1432	KIAA1432	-	-	-	membrane
ATXN3	ATXN3	ataxin 3	Protein processing in endoplasmic reticulum	cellular response to misfolded protein/ response to abiotic stimulus	cysteine-type peptidase activity	nuclear inclusion body
TRPC4AP	TRPC4AP	transient receptor potential cation channel, subfamily C, member 4 associated	Protein modification	ubiquitin-dependent protein catabolic process	receptor activity	Cul4A-RING ubiquitin ligase complex
NHS	NHS	Nance-Horan Syndrome	-	cell differentiation	-	tight junction

[a]**Hs Symbol**: *Homo Sapiens* gene symbol.

Once the regulatory sequence analysis performed, we identified 25 transcription factors (TF) strongly related (p-value<0.001) with 10 of 13 genes identified as shown in the supplementary table (ESM_3.pdf). The top three TF were choosing for construction of a network with their pathways and gene ontology (Fig. 2b).

Discussion

Statistical Analyses

We used a Bayesian multiple regression models considering different distribution for the data (Gaussian and Poisson). Once the number of teats (NT) is considered as a count phenotype, we checked the efficiency of Gaussian against Poisson distribution in a SNPs association study. In the analysis, the deviance value was significantly smaller when using the Gaussian model (Fig. 1), since the 95% HPD interval for the deviance difference (Gaussian less Poisson) did not contain the value zero, containing only negative values. These results indicate that, based on this criterion, this model is the most appropriate for estimating the marker effects for the number of teats by presenting better fitting to the observed data and also a lower degree of complexity. A special comment must be given for the estimated SNP covariance matrix, since we observed significant covariance between SNPs, i.e., for a large number of SNPs pairs, the HPD intervals did not include the zero value for the covariance. It is relevant due the most GWAS analysis including all SNPs simultaneously in the models (e.g. Bayes CPi and Bayes DPi), its effects are considered independents. Thus, the present study presents a practical and general way to take into account this dependence by assuming a covariance matrix based on previous LD analysis.

Although the phenotype used is characterized as a counting discrete variable, the Gaussian distribution better described the behavior of this trait when compared to the Poisson distribution. This can be explained by the fact that the Poisson distribution is asymmetric and skewed right, and even though it is continuous, the symmetry of the

Gaussian distribution ensured the best fit, most likely because it was more consistent with the observed distribution of sample data. Another possible explanation is that the Poisson distribution assumes that the variable's mean is equal to its variance, a condition that may not have been met when using the study data. Thus, in future studies, it would be interesting to consider other distributions for discrete random variables for which such an assumption would not need to be met, such as the negative binomial distribution.

Regarding the polygenic effect in the model, we observed no improvement in the fitting quality given the similarity between the deviance posterior distribution from models with and without this effect (ESM_2). Similar results were cited by Silva et al. (2011), which studying this same population did not observed significant gains in the QTL detections for carcass traits when including the polygenic effect in the evaluated models. In general terms, since the SNPs are located at known QTL regions that explain a large amount of genetic variance, and also due to F2 structure that provides a certain kind of homogeneity in relationship coefficients, the inclusion of the polygenic effect in the GWAS models was not significant. Nevertheless, this effect must be tested in GWAS analysis because in some populations, even in the presence of high density SNP panels, it can be used to adjust for population structure differences.

Considering the results from the best model (Gaussian distribution), a total of 19 significant SNPs, distributed in chromosomes (SSC) 1, 4, 7, 8, 17 and X (Table 1) were indentified. This significance was accessed by 95% HPD intervals, such that the value zero is not included in the interval for the marker effect the SNP in question was declared as significant at a probability level of 5%. Even results of GWAS for number of teats being scarce, several QTL for that trait were previously reported in the same chromosome regions identified on this study. At SSC1 we identified SNPs overlapping QTL locations from studies for Meishan x Gottingen (Wada et al., 2000) and Meishan x

Large White cross population (Guo et al., 2008).In the Guo et al. study, they found also QTL for number of teats overlapping our markers location onSSC7 and SSC17.

At SSC4, Bidanel et al. (2008) reported a QTLfor NT in a Chinese Meishan x European Large White cross population. Ding et al. (2009) also reported a QTL in a White Duroc x Erhualian pig resource population for that chromosome. Beeckmann et al. (2003), in a study of Linkage and QTL mapping for SSC 8, reported a QTL overlapping our marker identified on this chromosome. Cepicaet al. (2003) in a Linkage and QTL mapping for SSC X reported a QTL overlapping our marker identified on this chromosome for number of teats. The identification of several new markers associated with teat number traits in this study and the confirmation of QTL identified in earlier experiments can help us to evaluate the different markers effects on teat numbers and their biological function when added with post GWAS analyses as genes networks.

Genes Network Analysis

Among the 19 significant SNPs, 13 genes were identified and grouped into a network of functional relevance. They were grouped by features in common between them (ie. component cellular as Nucleus withSH3BGRL2 and ICK). The SH3BGRL2 gene description is a SH3 domain-binding glutamic acid-rich-like protein 2-like. In human it is involved in the control of redox-dependent processes and interact with Protein kinase C-θ (PKCθ) resulting in the inhibition of transcription factors like c-Jun, AP-1, and NF-κB (Mazzocco et al., 2002). Protein interactions involving SH3 domains have been reported to been involved in signal transduction, cytoskeleton rearrangements, membrane trafficking, and other key cellular processes (Cesareniet al., 2002). Here this gene has a polymorphism in it sequence related to number of teats trait and it is sharing the same component cellular with ICK (intestinal cell kinase) gene. The intestinal cell kinase gene has been better studied in human and may be involved in cell-

cycle regulation and apoptosis during mammalian development, suggesting that ICK plays a key role in the development of multiple organ systems (Lahiry et. al., 2009).

We identified also genes sharing the same biological process as response to abiotic stimulus (ATXN3 and GRM1) and genes with molecular function in common as cysteine-type peptidase activity (ATXN3 and TINAGL1). The GRM1 gene is a Metabotropic Glutamate 1Receptors. Studies in human identified its presence in many brain structures as the olfactory circuitry, hypothalamus (Shigemoto et al., 1992), basal ganglia (Testa et al., 1994) and it is related to breast cancer (Mehta et al., 2013). Sharing this biological process with this gene we had the ATXN3. This gene is also known as Machado-Joseph disease (MJD) and encode for Ataxin-3 protein which might play roles in neurodegeneration and modulate the aggregation of abnormal peptides in the pathogenesis of the diseases in human (Chen et. al., 2012). It has been reported to have it expression altered when induced by stradiol, diethylstilbestrol and octyl-phenol in the uterus of immature rats (Hong et. al., 2006).

The ATXN3 also is sharing the cysteine-type peptidase activity molecular function with TINAGL1 gene on the network analyzed here. This gene is also known as tubulointerstitial nephritis antigen-like 1and was cited to be differentially expressed in epithelial teat tissue of pigs in studies of QTL region–specific of positional candidate genes associated with the inverted teat defect (Chomwisarutkun et. al., 2013). The other genes of the network analyzed here did not have features in common each other but they entered at the net to be related with the trait analyzed.

Besides, we explored the promoter regions of the genes predicted to be targeted by the top TFs, NR3C1, NFIL3 and SP1, to construct a network of TF-target. NR3C1 is a Nuclear Receptor Subfamily 3, Group C, Member 1which encodes a glucocorticoid receptor (GR). Studies suggest that GRregulates mammary epithelial cell proliferation during late lobuloalveolar development (Wintermantel et al., 2005). Here this TF was

associated with seven genes (GRM1, ATXN3, SH3BGRL2, ICK, KIAA1432, CSGALNACT1 and NHS). NFIL3 is a Nuclear Factor, Interleukin 3 Regulated gene also known as E4BP4. Cowell (2002) demonstrated that E4BP4 has a widerange of physiological functions working in concert with members of the PAR family of transcription factors (e.g., on the regulation of apoptosis). Here this TF was associated with the following genes: GRM1, SH3BGRL2, ICK, KIAA1432, ATXN3, TRPC4AP and NHS.

The third TF SP1 is a specificity protein which was the first transcription factor identified and cloned (Dynan and Tjian, 1983). There is evidence that SP proteins may play a role in the growth and metastasis of many tumor types, including breast, by regulating expression of cell cycle genes and vascular endothelial growth factor (Safe and Abdelrahin, 2005). Here this TF was the most representative associated in a network with eight SNP genes (GRM1, ATXN3, TRPC4AP, SH3BGRL2, IER5L, TINAGL1, KIAA1432 and NHS), it biological process was in utero embryonic development and it pathway in development Prolactin receptor signaling.

Two genes (TRPC4AP and IER5L) which did not shared any ontology in the previously network but were highly related with the three TF and have evidences to be associated with human breast. The TRPC4AP gene is a transient receptor potential cation channel, subfamily C, member 4 associated protein. This gene has been cited to be down regulated in a study which compared the gene expression profiles in normal breast epithelia from parous postmenopausal women with and without breast cancer (Balogh et al., 2007). The IER5L gene is an immediate early response 5-like protein and it expression is cited to be downregulated by Arsenic trioxide in women MCF-7 breast cancer cells(Wang et al., 2011).

Concluding Remarks

Our comparative statistical analyses allowed us to confirm the superiority of Gaussian in relation to Poisson distribution for the trait number of teats. Although Poisson is appropriated to count data, its assumption with respect to the equivalence of mean and variance sometimes can to impair its fitting to biological data. Thus, more powerful distributions as Negative Binomial and Generalized Poisson can be used as alternative distributions for counting phenotypes. However, firstly it is necessary the development of computational tools that contemplate them in GWAS models.

The F2 populations have a great power to detect QTL provided by linkage disequilibrium, but also make it difficult to discriminate between causal and neutral mutations. In this context, more sophisticated models initially proposed to QTL detection especially in F2 populations can be adapted to GWAS analysis. Among these models, stand out those proposed by Varona et al. (2005), which include simultaneously the genetic configuration of the mutation and the probability of line origin given the neutral markers. Thus, in future researches generalizations of this model can be proposed in order to point out for non-normal phenotypes and SNP effect estimation.

The present study provided a rich information resource about genes related to the number of teats in pigs, increasing our understanding of the molecular mechanisms underlying them. The genes network analysis predicted interactions that were consistent with the known mammal's breast biology and captured known regulation binding sites, allowing the identification of new candidate genes(e.g., TINAGL1 and ICK). However, the number of teats is a complex trait that is subject to the action ofa large number of genes that are regulated by several transcription factors, therefore many of them still to be identified.

Acknowledgements

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado de Minas Gerais

(FAPEMIG) and Coordenção de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/NUFFIC and CAPES/DGU).

Ethical standards

The use of animals was reviewed and approved by the Ethics Statements of the Department of Animal Science, Federal University of Viçosa (UFV), MG, Brazil.

Reference

Ayres DR, Pereira RJ, Boligon AA, Silva FF, Schenkel FS, Roso VM, Albuquerque LG (2013) Linear and Poisson models for genetic evaluation of tick resistance in cross-bred Hereford x Nellore cattle. J. Anim. Breed. Genet., DOI: 10.1111/jbg.12036.

Balogh GA, Russo J, Mailo DA, Heulings R, Russo PA, Morrison P,Sheriff F, Russo, IH (2007) The breast of parous women without cancer has a different genomic profile compared to those with cancer. Int. J.oncol., 31(5):1165.

Band GO, Guimarães SEF, Lopes PS, Schierholt AS, Silva KM, Pires AV, Júnior AAB, Gomide LAM (2005a) Relationship between the Porcine Stress Syndrome gene and pork quality traits of F_2 pigs resulting from divergent crosses. Genet. Mol. Biol., 28:88-91.

Band GO, Guimarães SEF, Lopes PS, Peixoto JDO, Faria DA, Pires AV, Figueiredo FC, Nascimento CS, Gomide LAM (2005b) Relationship between the Porcine Stress Syndrome gene and carcass and performance traits of F_2 pigs resulting from divergent crosses. Genet. Mol. Biol., 28:92-96.

Beeckmann P, Moser G, Bartenschlager H, Reiner G, Geldermann H (2003) Linkage and QTL mapping for Susscrofa chromosome 8. J. Anim. Breed. Genet., 120(1); 66-73.

Bidanel JP, Rosendo A, Iannuccelli N, Riquet J, Gilbert H, Caritez JC, Billon Y, Amigues Y, Prunier A, Milan D (2008) Detection of quantitative trait loci for teat number and female reproductive traits in Meishan x Large White F2 pigs. Anim., 2(6):813-820

Cepica S, Reiner G, Bartenschlager H, Moser G, Geldermann H (2003) Linkage and QTL mapping for Sus scrofa chromosome X. J Anim Breed Genet., 120(1): 144-151

Cesareni G, Panni S, Nardelli G, Castagnoli L (2002) Can we infer peptide recognition specificity mediated by SH3 domains? Federation Eur. Biochem. Societies, 513:38–44.

Clayton GA, Powell JC, Hiley PG (1981). Inheritance of teat number and teat inversion in pigs. Anim. Prod. 33:299–304.

Chen CP, Chen YH, Chern SR, Chang SJ, Tsai TL, Li SH, Chou HC, Lo YW, Lyu PC, Chan HL (2012) Placenta proteome analysis from Down syndrome pregnancies for biomarker discovery. Mol. BioSyst., 8:2360–2372.

Chomwisarutkun K, Murani E, Brunner R, Ponsuksili S, Wimmers K (2013) QTL regions-
ecific microarrays reveal differential expression of positional candidate genes of sig naling pathways associated with the liability for the inverted teat defect. Anim. Genet.,44(2):139–148.

Cowell IG (2002) E4BP4/NFIL3, a PAR-related bZIP factor with many
roles. Bioessays, 24(11):1023-1029.

Cui Y, Kim D-Y, Zhu J (2006) On the Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci With Count Data. Genet., 174 (4):2159–2172.

Ding N, Guo Y, Knorr C, Ma J, Mao H, Lan L, Xiao S, Ai H, Haley CS (2009) Genome-wide QTL mapping for three traits related to teat number in a White Duroc x Erhualian pig resource population. BMC Genet, 10:6.

Dynan WS, Tjian R (1983) The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell, 35(1):79-87.

Fortes MR, Reverter A, Zhang Y, Collis E, NagarajSH, Jonsson NN, PrayagaKC, Barris W,Hawken RJ (2010) Associationweight matrix for the genetic dissection of puberty in beef cattle. Proc. Natl. Acad. Sci. USA, 107:13642–13647.

Fortes MR, Reverter-Gomez T, Hiriyur-Nagaraj S, Zhang Y, Jonsson N, Barris W, Lehnert S, Boe-Hansen GB,Hawken R (2011) A SNP-derived regulatory gene network underlying puberty in two tropical breeds of beef cattle. J. Anim. Sci., 89:**1669**–1683.

Friedrichs N, Steiner S, Buettner R, Knoepfle G (2007) Immunohistochemical expression patterns of AP2α and AP2γ in the developing fetal human breast. Histopathol., 51(6):814–823.

Guo Y-M, Lee GJ, Archibald AL, Haley CS (2008) Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan x Large White populations. Anim.Genet., 39 (5):486-95.

Habier D, Fernando RL,DekkersJCM (2009) Genomic Selection Using Low-Density Marker Panels. Genet. 182:343–353.

Harville DA, Callanan TP (1989) Computational aspects of likelihood-based inference for variance components. In: Gianola D, Hammond K (ed) Advances in Statistical Methods for Genetic Improvement of Livestock, edn. Springer-Verlag, Berlin, pp 136–176

Hidalgo AM, Lopes PS, Paixão DM, Silva FF, BastiaansenJWM, Paiva SR, Faria DA, Guimarães SEF (2013) Fine mapping and single nucleotide polymorphism effects estimation on pig chromosomes 1, 4, 7, 8, 17 and X. Genet. Mol. Biol., 36(4): 511-519.

Hirooka H, de Koning DJ, Harlizius B, van Arendonk JA, Rattink AP, Groenen MA, Brascamp EW, Bovenhuis H (2001) A whole-genome scanfor quantitative trait loci affecting teat number in pigs.J. Anim. Sci., 79(9):2320-2326.

Hong E-J, Park S-H, Choi K-C, Leung PCK, Jeung E-B (2006) Identification of estrogen-regulated genes by microarray analysis of the uterus of immature rats exposed to endocrine disrupting chemicals. Reproduc. Biol.Endocrinol., 4:49.

Lahiry P, Wang J, Robinson JF, Turowec JP, et al. (2009) A Multiplex Human Syndrome Implicates a Key Role for Intestinal Cell Kinase in Development of Central Nervous, Skeletal, and Endocrine Systems. Am.J. Hum. Genet., 84(2):134–147.

Liu Y, Lee YF, Ng MK (2011) SNP and gene networks construction and analysis from classification of copy number variations data. BMC Bioinform., 12(Suppl 5):S4.

Lorenzo-Bermejo J, Beckmann L, Chang-Claude J, Fischer C (2011) Using the posterior distribution of deviance to measure evidence of association for rare susceptibility variants.BMC Proc., 5(9):S38.

McKay RM, Rahnefeld GW (1990) Heritability of teat number in swine. Can. J. Anim. Sci. 70:425–430.

Martin AD, Quinn KM, Park JH (2011)MCMCpack: Markov chain Monte Carlo in R. J. Stat.Softw., 42(9):1-21.

Mazzocco M, Maffei M, Egeo A, Vergano A, Arrigo P, Di LR, Ghiotto F, Scartezzini P (2002) The identification of a novel human homologue of the SH3 binding glutamic acid-rich (SH3BGR) gene establishes a new family of highly conserved small proteins related to thioredoxin superfamily. Gene, 291:233–239.

Mehta MS, Dolfi SC, Bronfenbrener R, Bilal E, Chen C, Moore D et al. (2013) Metabotropic Glutamate Receptor 1 Expression and Its Polymorphic Variants Associate with Breast Cancer Phenotypes. .PloSone, 8(7):e69851.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genet.,157(4):1819–1829.

Onteru SK, Fan B, DuZ-Q, Garrick DJ,StalderKJ, RothschildMF(2011)A whole-genome association study for pig reproductive traits. Anim. Genet., 43:18–26.

Perez-Enciso M, Tempelman RJ,Gianola D (1993)A comparisonbetween linear and Poisson mixed models for litter size inIberian Pigs. Livest. Prod. Sci.,35:303.

R Development Core Team (2012) R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE et al. (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE, 4:e6524.

Reverter A,FortesMRS (2013) Building single nucleotide polymorphism-derived gene regulatory networks: Towards functional genomewide association studies. J. Anim. Sci., 91:530-536.

Safe S,Abdelrahim M (2005) Sp transcription factor family and its role in cancer. Eur. J. Cancer, 41(16):2438-2448.

Schneider JF,RempelLA, Rohrer GA (2012) Genome-wide association study of swine farrowing traits. Part I: Genetic and genomic parameter estimates. J. Anim. Sci., 90:3353-3359.

Shigemoto R, Nakanishi S, Mizuno N (1992) Distribution of the mRNA for a metabotropic glutamate receptor (mGluR1) in the central nervous system: an in situ hybridization study in adult and developing rat. J. Comp. Neur., 322**:**121-135.

Silva KM, Knol EF, Merks JWM, Guimarães SEF, Bastiaansen JWM, van Arendonk JAM, Lopes PS (2011) Meta-analysis of results from quantitative trait loci mapping studies on pig chromosome 4. Anim. Genet., 42:280-292.

Silva FF, Rosa GJ, Guimarães SE, Lopes PS, de los Campos G (2011) Three-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. Livest. Sci., 142(1), 210-215.

Testa CM, Standaert DG, Young AB, Penney JB Jr. (1994) Metabotropic glutamate receptor mRNA expression in the basal ganglia of the rat. J.Neurosci.,14**:**3005-3018.

Touzet H, Varré JS (2007) Efficient and accurate P-value computation for Position Weight Matrices. Algorithms Mol. Biol., 2:15.

Uimari P,Sironen A,Sevón-Aimonen M-L (2011)Whole-genome SNP association analysis ofreproduction traits in the Finnish Landrace pigbreed. Genet. Sel.Evol., 43:42.

Varona L, Gómez-Raya L, Rauw WM, Noguera JL(2005)A simulation study on the detection of causal mutations from F2 Experiments. J. Anim. Breed.Genet. 122:30–36

Vazquez AI, Bates DM, Rosa GJM, Gianola D, Weigel KA (2010) Technical note: An R package for fitting generalized linear mixed models in animal breeding. J. Anim. Sci., 88(2), 497-504.

Varona L, Sorensen D (2010) A genetic analysis of mortality in pigs. Genet., 184:277.

Wada Y, Akita T, Awata T, Furukawa T, Sugai N, Inage Y, Ishii K, Ito Y,Kobayashi E, Kusumoto H, Matsumoto T, Mikawa S, Miyake M (2000) Quantitative trait loci (QTL) analysis in a Meishan x Gottingen crosspopulation. Anim. Genet., 31(6):376-84.

Wang X, Gao P, Long M, Lin F, Wei JX, Ren JH et al. (2011) Essential role of cell cycle regulatory genes p21 and p27 expression in inhibition of breast cancer cells by arsenic trioxide. Medical Oncol.,28(4):1225-1254.

Wintermantel TM, Bock D, Fleig V, Greiner EF, Schütz G (2005) The epithelial glucocorticoid receptor is required for the normal timing of cell proliferation during

mammary lobuloalveolar development but is dispensable for milk production. Mol.Endocrinol., 19(2): 340-349.

Zheng X, Levine D, Shen J,Gogarten SM, Laurie C, Weir BS (2012) A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. Bioinform.,28(24):3326-3328.

Supplementary data

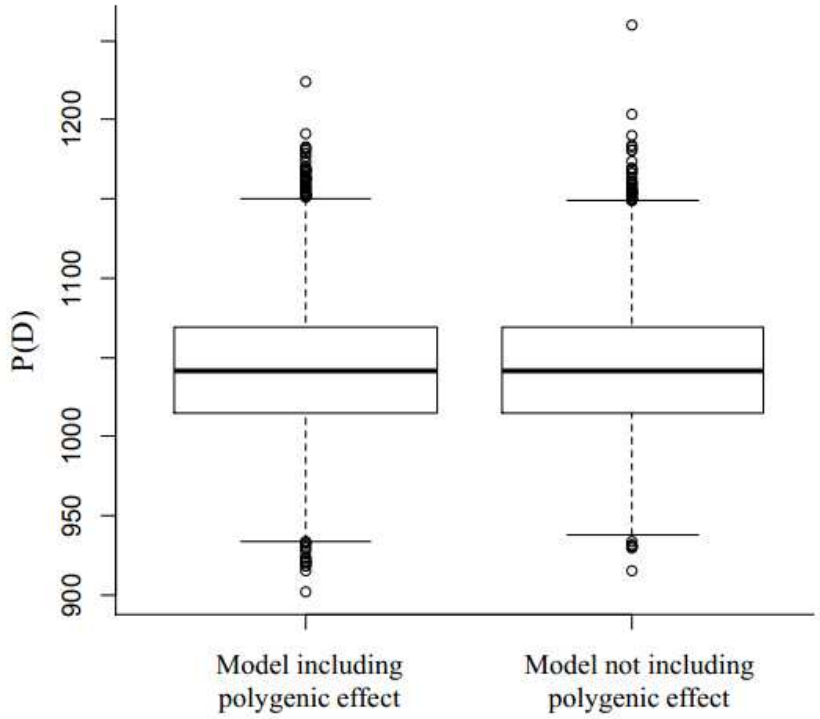


Figure ESM_2 The box plot of deviance posterior distribution P(D) for GWAS models fitted to number of teats in pigs including, or not, the polygenic effect.

Table ESM_3 The 25 transcription factors (TF) clusters strongly related (P-value<0.001) with identified genes (Target Genes) for number of teats obtained from TFM-explorer tool.

TF	Target Genes[a]	P-value
NR3C1	GRM1, SH3BGRL2, ICK, KIAA1432, ATXN3, CSGALNACT1 and NHS	0.00001
NFIL3	GRM1, SH3BGRL2, ICK, KIAA1432, ATXN3, TRPC4AP and NHS	0.00003
SP1	GRM1, SH3BGRL2, IER5L, TINAGL1, KIAA1432, ATXN3, TRPC4AP and NHS	0.00003
FOXI1	GRM1, IER5L, TINAGL1, ICK, KIAA1432, TRPC4AP, CSGALNACT1 and NHS	0.00003
NFE2L1::MafG	GRM1, IER5L and CSGALNACT1	0.00005
TBP	GRM1, SH3BGRL2, TINAGL1, ATXN3, TRPC4AP, CSGALNACT1 and NHS	0.00005
TFAP2A	SH3BGRL2, IER5L, TINAGL1, ICK, KIAA1432, ATXN3 and TRPC4AP	0.00008
FOXO3	GRM1, TINAGL1, ICK and ATXN3	0.0002
NFATC2	GRM1, SH3BGRL2, IER5L, ICK, KIAA1432, TRPC4AP, CSGALNACT1 and NHS	0.0002
NR2F1	SH3BGRL2, TINAGL1, ICK, ATXN3, CSGALNACT1 and NHS	0.0003
FOXD1	GRM1, SH3BGRL2, IER5L, ICK, KIAA1432, ATXN3, TRPC4AP and NHS	0.0003
Foxq1	GRM1, SH3BGRL2, TINAGL1, ICK, KIAA1432, ATXN3, TRPC4AP, CSGALNACT1 and NHS	0.0003
MYC::MAX	KIAA1432, CSGALNACT1 and NHS	0.0004
MEF2A	GRM1, SH3BGRL2, ICK, KIAA1432, ATXN3, TRPC4AP, CSGALNACT1 and NHS	0.0004
TBP	IER5L, ICK, ATXN3 and NHS	0.0004
Mycn	SH3BGRL2, IER5L, KIAA1432, ATXN3, TRPC4AP and NHS	0.0004
USF1	GRM1, IER5L, TINAGL1, KIAA1432, ATXN3, TRPC4AP, CSGALNACT1 and NHS	0.0004
MAX	GRM1, TINAGL1, ICK, KIAA1432, CSGALNACT1 and NHS	0.0005
Zfx	IER5L, TINAGL1, KIAA1432, ATXN3, CSGALNACT1 and NHS	0.0006
Sox17	SH3BGRL2, IER5L, TINAGL1, ATXN3, TRPC4AP and CSGALNACT1	0.0006
RXR::RAR_DR5	SH3BGRL2, TINAGL1, ICK, KIAA1432, ATXN3, CSGALNACT1 and NHS	0.0007
Hand1::Tcfe2a	GRM1, SH3BGRL2, TINAGL1, ICK, TRPC4AP, CSGALNACT1 and NHS	0.0007
Klf4	IER5L, TINAGL1 and ATXN3	0.0008
FEV	GRM1, SH3BGRL2, TINAGL1 and ICK	0.0008
PLAG1	SH3BGRL2, IER5L, TINAGL1, ICK, KIAA1432, ATXN3, TRPC4AP and NHS	0.0008

[a] GRM1: glutamate receptor, metabotropic 1; SH3BGRL2: SH3 domain-binding glutamic acid-rich-like protein 2-like; IER5L: immediate early response gene 5-like protein-like; TINAGL1: tubulointerstitial nephritis antigen-like; ICK: intestinal cell (MAK-like) kinase; KIAA1432: KIAA1432; ATXN3: ataxin 3 and TRPC4AP : transient receptor potential cation channel, subfamily C, member 4 associated; CSGALNACT1: Chondroitin Sulfate N-Acetylgalactosaminyltransferase 1 and NHS: Nance-Horan Syndrome.

Chapter 3

Revealing new candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways

Lucas L Verardo[1, 2*], Fabyano F Silva[1], Marcos S Lopes[2, 3], Ole Madsen[2], John WM Bastiaansen[2], Egbert F Knol[3], Mathew Kelly[4], Luis Varona[5], Paulo S Lopes[1] and Simone EF Guimarães[1]

[1]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, 36570000, Brazil
[2]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6700 AH, The Netherlands
[3]Topigs Norsvin, Research Center, Beuningen, 6641 SZ, the Netherlands
[4]Queensland Alliance for Agriculture & Food Innovation, The University of Queensland, Queensland, QLD 4072, Australia
[5]Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, Zaragoza, 50013, Spain

Abstract

Background

Reproductive traits such as stillborn (SB) and number of teats (NT) have been evaluated in many genome wide association studies (GWAS). Most of these GWAS were performed assuming that these traits are normally distributed, however both SB and NT are discrete (e.g. count) variables. Therefore, it is necessary to test for better fit of other appropriate statistical models based on discrete distributions. In addition, even though many GWAS have been performed, only a few studies have explored biological meanings of genes identified. Applying gene networks, GWAS results can be used for a better genetic dissection of complex phenotypes. In this study, we performed and tested a Bayesian treatment of a GWAS model assuming Poisson distribution for SB and NT in a commercial pig line. Moreover, we used the significant SNPs to obtain the related genes and generate gene-transcription factor networks aiming at exploring biological roles behind the considered traits and pointing out the most probable candidate genes.

Results

Poisson and Gaussian distribution comparisons analysis showed that the Poisson model was appropriate for SB, whilst the Gaussian was appropriate for NT. The fitted GWAS models indicated 18 and 65 significant SNPs with one and nine QTL blocks, respectively for SB and NT. Within these results, we identified 18 genes related to SB and 57 to NT; these genes then were used to identify related transcription factors (TF). From these, we selected the most representative TF for each trait and constructed a gene-TF network illustrating the gene-gene interaction, and pointing out new candidate genes.

Conclusions

Our comparative analyses showed a significant better fit of Bayesian GWAS Poisson model for SB. Thus, counting models should be considered when analysing count traits aiming to increase analyses accuracy. Candidate genes (e.g. *PTP4A2, NPHP1* and *CYP24A1* for SB and *YLPM1, SYNDIG1L, TGFB3* and *VRTN* for NT) and TF (e.g. *NF-kappaB* and *KLF4* for SB and *SOX9* and *ELF5* for NT) were found. These results were coherent with known newborn survival traits (e.g. congenital heart disease in fetuses; kidney diseases and diabetes in the mother) and breast biology (e.g. mammary gland development and body length).

Background

Reproductive traits, such as stillborn piglets (SB) and number of teats (NT), are widely included in the selection indices of pig breeding programs due their importance to the pig industry. The number of stillborn piglets is a complex trait that is directly affected by the total number born [1] and temporal gene effects in different parities [2]. In humans, it has been shown that kidney diseases and diabetes in the mother and congenital heart disease in the fetuses are some of the main cause of its occurrence [3]-[5]. In pigs, however, a better biological understanding of these traits is still needed to improve selection against SB in pig breeding. Number of teats (NT) is a trait with a large influence on the mothering ability of sows [6], being a limiting factor for the increased number of weaned piglets. Biologically, the development of embryonic mammary glands require the coordination of many signaling pathways to direct the cell shape changes, cell movements, and cell–cell interactions necessary for proper morphogenesis of mammary glands [7]. In addition, number of vertebrae, that will determine the body length of the sow, may also have a direct relation with the final NT observed in pigs [8].

As these traits are directly involved with higher production and welfare of piglets, several genome-wide association studies (GWAS) have been performed for SB and NT [2], [9], [10]. However, in these GWAS, these traits were presumed normally distributed, which may not be true. Both, SB and NT, are measured as count variables, and therefore they follow discrete distributions such as Poisson. Although Poisson distribution has already been implemented in animal breeding in the context of traditional mixed models [11], [12] and Quantitative Trait Loci (QTL) mapping [13], [14], there are no reports of GWAS for SB and NT using this approach.

Poisson models can be performed using a Bayesian Markov chain Monte Carlo (MCMC) approach [15]. Applying Poisson models to GWAS, using a traditional mixed model, it is possible to assume all markers simultaneously by using the genomic relationship matrix [16], as preconized in genomic best linear unbiased prediction (GBLUP). GBLUP has been widely used in genome wide selection (GWS) on which the GEBV vector, from each MCMC iteration, can be directly converted into SNP marker effects vector [17], [18]. Based on this, samples of posterior distribution for each particular SNP effect are generated at each cycle and significance tests, based on Highest Posterior Density (HPD) intervals, can be performed in order to identify the most relevant markers. Studies have also adopted the HPD intervals to infer directly on marker effects significance [19], [20]. Additionally, the posterior probability (PPN0) of the estimated effect being lower than 0 (for negative effects) or greater than 0 (for positive effects) as proposed by Ramírez et al. [20] and Cecchinato et al. [21] can be also used to report marker effect significance under a Bayesian approach. Furthermore, when considering all markers simultaneously in the model, some current issues like the influence of gene length [22] can be minimized by estimating its particular effects in the presence of all others markers effects.

Although many GWAS have been performed, only a few have explored the biological meaning of the identified genes. GWAS results can be used for the genetic dissection of complex phenotypes through applying a network approach to the genes identified in the genomic regions around significant SNPs. The genes in linkage to significant SNPs can be used to examine the sharing of pathways and functions as well as the enrichment of related transcription factors (TF) significantly in the selected genes. TFs have been shown to associate with important traits in pigs e.g. *PIT1* for carcass traits [23] and *SREBF1* in regulation of muscle fat deposition [24]. Thus, providing evidence for an interaction between a known trait-related TF and its predicted target genes via regulatory sequence analysis and gene-TF network constructions, serves as an *in silico* validation for the gene-gene interactions. The gene-TF network will facilitate the identification of the most probable group of candidate genes for the studied traits. Similar approaches have been performed for puberty related traits in cattle [25], [26] identifying related candidate genes and transcription factors. In pigs, however, this approach has just begun to be more exploited [27].

In this study, we performed a Bayesian treatment of a GWAS model assuming Poisson and Gaussian distribution for SB and NT in a commercial pig line. Moreover, we used the significant SNPs to obtain the related genes and generate gene-transcription factor networks aiming at exploring biological roles behind the considered traits and pointing out the most probable candidate genes.

Methods

Phenotypic and Genotypic Data

Stillborn records from 1,390 Large White (LW) sows with an average of 3.9 parities were evaluated. The average SB number in this population was 1.2, ranging from 0 to 16 stillborn piglets. NT was counted at birth for a total of 1,795 LW animals. The

average NT in this population was 15.3, ranging from 14 to 20 teats. All animals were genotyped using the Illumina 60K+SNP Porcine Beadchip [28]. As part of quality control procedures, SNPs with GenCall <0.15 [29], minor allele frequency <0.01, frequency of missing genotypes >0.05, unmapped SNPs and SNPs located on Y chromosome, according to the Sscrofa10.2 assembly of the reference genome [30] were excluded from the data set. SNPs showing a genotype call rate <0.95 were also excluded. After quality control, genotypes of 1,657 (NT) and 1,200 (SB) animals for 41,647 SNPs were included in the association analyses.

Statistical Analysis

Two genomic best linear unbiased predictor (GBLUP) models were fitted to the data under a Bayesian framework. The difference between these models was that one assumed that the traits follow a Gaussian distribution and the second one assumed that the traits follow a Poisson distribution. For the Gaussian response, the following general linear model was assumed:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e} \quad (1)$$

Where y was the vector of phenotypic observations; β was the vector of systematic environmental effects; u was a vector of additive genetic effects, p was a vector of permanent environment effects (fitted only in the model for SB) and e was a vector of residual effects; X, Z and W were design matrices related to β, u and p, respectively. The herd-year-season (HYS) was considered as systematic effect for both SB and NT, while a sow parity number was used only for SB and a sex effect was used only for NT. Sex effects are commonly accounted when analyzing number of teats according Lopes et al. [31] who showed that males presented 0.35±0.09 more teats than females.

For the Poisson response **y**, a latent variable (**l**) was introduced by means of the canonical parameter **λ** (often called the rate or mean parameter) of this distribution and the link function on the log scale, i.e., $y_i \sim \text{Poi}(\lambda = \exp(\mathbf{l}_i))$, where exp is the inverse

link function. In this case, the linear model presented in (1) can be rewritten in terms of latent variable as follow:

$$\mathbf{l} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e} \quad (2)$$

In the Bayesian analysis, the following prior distributions were assumed for the parameters of models (1) and (2): $\boldsymbol{\beta}\sim\mathrm{N}(0, \sigma^2_{\beta}\mathbf{I})$, being σ^2_{β} known and assumed as large (in this case 1e+10) in order to represent vague prior knowledge; $\mathbf{u}\sim\mathrm{N}(0, \sigma^2_{\mathrm{u}}\mathbf{G})$ and $\mathrm{p}\sim\mathrm{N}(0, \sigma^2_{\mathrm{p}}\mathbf{I})$, with $\mathbf{I}$ being an identity matrix and $\mathbf{G}$ the genomic relationship matrix proposed by Van Raden [17]. Thus, $\mathbf{G} = \mathbf{MM}'/\Sigma^{\mathrm{J}}_{\mathrm{j}=1}2\mathrm{p_j}(1-\mathrm{p_j})$, where $\mathbf{M}$ was the incidence SNP matrix assuming 0, 1 and 2 for genotypes aa, aA and AA, respectively, and $\mathrm{p_j}$ was the allele frequency of SNP j. For the variance components (σ^2_{u}, σ^2_{p} and σ^2_{e}), an inverted chi-squared distribution was considered: $\sigma^2_{\mathrm{u}}|\,\mathrm{V_u}$, $\mathrm{S_u}\sim\mathrm{V_uS_uX}^{-2}_{\mathrm{v_u}}$, $\sigma^2_{\mathrm{p}}|\,\mathrm{V_p}$, $\mathrm{S_p}\sim\mathrm{V_pS_pX}^{-2}_{\mathrm{v_p}}$ and $\sigma^2_{\mathrm{e}}|\,\mathrm{V_e}$, $\mathrm{S_e}\sim\mathrm{V_eS_eX}^{-2}_{\mathrm{v_e}}$, where V and S are hyperparameters.

In both models (1 and 2), the residual vector was assumed as $\mathbf{e}\sim\mathrm{N}(0, \sigma^2_{\mathrm{e}}\mathbf{I})$, implying a Gaussian likelihood function. In agreement with Hadfield and Nakagawa [15], the conditional probability of the latent variable is proportional to the product of two terms, the Poisson likelihood of the data given $\mathbf{l}$ and the Gaussian likelihood based on residual terms, i.e., $\mathrm{P}\left(\mathbf{l_i}\middle|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{p}, \sigma^2_{\mathrm{u}}, \sigma^2_{\mathrm{p}}, \sigma^2_{\mathrm{e}}\right) \propto \prod^{\mathrm{N}}_{\mathrm{i}=1}\mathrm{P_i}(\mathrm{y_i}|\mathbf{l_i})\mathrm{P}(\mathrm{e_i}|\sigma^2_{\mathrm{e}})$. Thus, in model (2) the vector of latent variable ($\mathbf{l}$) was considered also as unknown parameter, therefore not having a recognizable full conditional distribution, implying using the Metropolis–Hastings algorithm to generate samples of posterior distribution for $\mathbf{l}$. For the other unknown parameters ($\boldsymbol{\beta}, \mathbf{u}, \mathbf{p}, \sigma^2_{\mathrm{u}}, \sigma^2_{\mathrm{p}}, \sigma^2_{\mathrm{e}}$), given the closed form of the full conditional posterior distributions, the Gibbs sampler algorithm was used. For the standard linear mixed model (1) with a Gaussian response and identity link, $\mathrm{P}(\mathbf{l_i}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{p}, \sigma^2_{\mathrm{u}}, \sigma^2_{\mathrm{p}}, \sigma^2_{\mathrm{e}})$ is always unity and so the Metropolis-Hastings steps are not required.

Models (1) and (2) were fitted to data using an adaptation of MCMCglmm R package [32] with a total of 100,000 iterations assuming a burn-in period of 50,000 and sampling interval (thin) each 2nd iteration. The adaptation described above is related with the inverted **G** matrix that replaced the traditional relationship matrix (**A**) in the *ginverse* option of this package.

The model comparisons were realized by means of the posterior distribution of the deviance $P(D_M)$ provided by a particular M (i.e. Poisson or Gaussian) model. For the Gaussian model, each value of this distribution was obtained directly by $D_G^{(k)} = -2\log\left(\prod_{i=1}^{N} P\left(y_i\middle|\boldsymbol{\theta}^{(k)}\right)\right)$,

in which $\prod_{i=1}^{N} P(y_i|\boldsymbol{\theta}^{(k)})$ is the value for the likelihood function considering the set of parameter estimates ($\boldsymbol{\theta}^{(k)}$) at each MCMC iteration k. Similarly, for the Poisson model, the values came from $D_P^{(k)} = -2\log\left(\prod_{i=1}^{N} P\left(y_i\middle|\lambda_i^{(k)}\right)\right)$, being $\lambda_i^{(k)}$ the estimate of Poisson mean, i.e. the exponential of the latent variable $\log(\lambda_i)$ generated by Metropolis-Hastings algorithm. Thus, the random draws from posterior distributions of the deviance for both models, $P(D_G)$ and $P(D_P)$, were used to simulate the distribution of deviance difference [33] given by $P(D_{G-P})$. From the obtained distributions, it was possible to propose a hypothesis test based on HPD (Highest Posterior Density) interval for the deviance difference. Lower deviance values indicate better fitting model, therefore, if the interval contained only negative values, the Gaussian model was indicated as the best. On the other hand, an interval containing only positive values implies the best fit of the Poisson model.

One important statistical feature of the present study was the fact that the vector **u** (GEBV) was kept in each k MCMC iteration ($\mathbf{u}^{(k)}$) from models (1) and (2), which enabled to generate the vector of SNP effects (**α**) in each iteration using the following linear system: $\boldsymbol{\alpha}^{(k)} = (\mathbf{M}'\mathbf{M})^{-}\mathbf{M}'\mathbf{u}^{(k)}$ being **M** the matrix of SNPs genotypes as defined

earlier. Once a vector of SNP effects was generated in each iteration ($\boldsymbol{\alpha}^{(k)}$), it was possible to obtain a MCMC chain with length equal to 25,000 iterations (see burn-in and thin already earlier mentioned) for each SNP marker. Thus, after verifying the convergence of these chains by Geweke and Raftery and Lewis criteria (using, respectively, the functions *geweke.diag* and *raftery.diag* of coda R package) [34], a sample of posterior distribution for the effect of each SNP was obtained. From these distributions was possible to calculate the 95% HPD (Highest Posterior Density) intervals as presented by Li et al. [19] and the posterior probability (PPN0) of the marker effect to be over (for positive effects) or below (for negative effects) zero as presented by Ramírez et al. [20] and Cecchinato et al. [21]. The HPD intervals and PPN0 were obtained for each marker, so that the chromosome positions of the significant SNPs were used for identifying genes influencing the target traits.

When obtaining the SNP effect directly by $(\mathbf{M}'\mathbf{M})^{-}\mathbf{M}$, as well as the storage of SNP´s chains, we faced a big size of data. Thus, this calculation was performed using the Intel Math Kernel Library, which is highly optimized for use on Intel processers and utilizes parallel process to decrease computational time. The application of these optimized libraries for matrix computation in genomics problems was discussed in detail in Aguilar et al [35]. The computer used to perform the analyses have 12 cores running Intel$^{(R)}$ Core$^{(TM)}$ i7-3930K CPU @ 3.20GHz and 96gb of ram.

Linkage disequilibrium (LD) relations between significant SNPs were evaluated through Haploview [36] to identify QTLs regions (blocks) among chromosomes using the default parameters based on Gabriel et al. [37]. A 95% confidence bounds on D prime [38] was used to determine if a pair of SNPs was in “strong LD”. Markers with MAF < 0.01 were ignored.

Gene-TF networks

Genes overlapping with significant SNP LD-blocks and individual significant non-LD SNPs including, for both, 32.5 Kbp flanking sequence (half average distance between SNPs presented on the chip) were identified at dbSNP NCBI web site (http://www.ncbi.nlm.nih.gov/SNP/). Checking this interval we verified the presence of any gene that could be related with a QTL or a significant SNP. Besides, studies on Large White breed have demonstrated that even for an average distance between two SNPs being around 200-250 Kb, the LD (r2) is still high (0.31) being an average LD greater than 0.2 reported to be required for genomic analyses [39], [40]. The overlapping genes were used to obtain Functional Gene Ontology (GO) terms and pathways of the genes with GeneCards (http://www.genecards.org/) and TOPPCLUSTER (http://toppcluster.cchmc.org/).

TF enriched in the identified set of genes were found with the TFM-Explorer program (http://bioinfo.lifl.fr/TFM/TFME/). This program takes a set of gene sequences, and searches for locally overrepresented transcription factor binding sites (TFBS) using weight matrices from JASPAR database [41] to detect all potential TFBS, and extracts significant clusters (region of the input sequences associated with a factor) by calculating a score function. This score threshold is chosen to give a *P*-value equal or better to 10^{-3} for each position for each sequence such as described in Touzet and Varré [42]. From the set of genes, excluding ncRNA genes, we collected 3000 bp upstream and 300 bp downstream of the gene start site based on Sscrofa10.2 assembly at NCBI web site. This data was used as input at TFM-explorer and the given list of TF was feed into Cytoscape [43] using a Biological Networks Gene Ontology tool (BiNGO) plugin [44], to determine the GO terms significantly overrepresented. Based on biological process overrepresented at BiNGO and evidence from literature review, we were able to identify the most representative TF related to our traits to construct a network with the interactions (gene-TF). Aiming to point out the most probable candidate genes for the

studied traits, we used the network analyses Cytoscape tool based on the number of TFBS and consequently number of connections in each gene to determine which are the most connected with our traits. Based on that, genes with more TFBS, for the most representative TF, must have more edges being highlighted at gene-TF network.

Results

Statistical Analyses

The most appropriate model was determined (Poisson or Gaussian distributions) for each trait by using the posterior distribution of deviance difference ($P(D_{G-P})$). The deviance was significantly smaller using the Poisson model (Figure 1a) for the SB trait, with the 95% HPD interval limits for $P(D_{G-P})$ ranging between 6614.86 (lower) and 7110.74 (upper). This interval does not contain zero values and only includes positive values, which are directly related with the highest deviance value for a Gaussian model suggesting a significantly better fit of a Poisson model for SB. On the other hand, for the trait NT, the Gaussian model presented significant lower deviance (Figure 1b), with the lower and upper HPD limits between -4300.66 and -3987.29, respectively. Thus, since the difference was computed as deviance from Gaussian model minus the deviance from Poisson model, the Gaussian model was clearly superior for this trait.

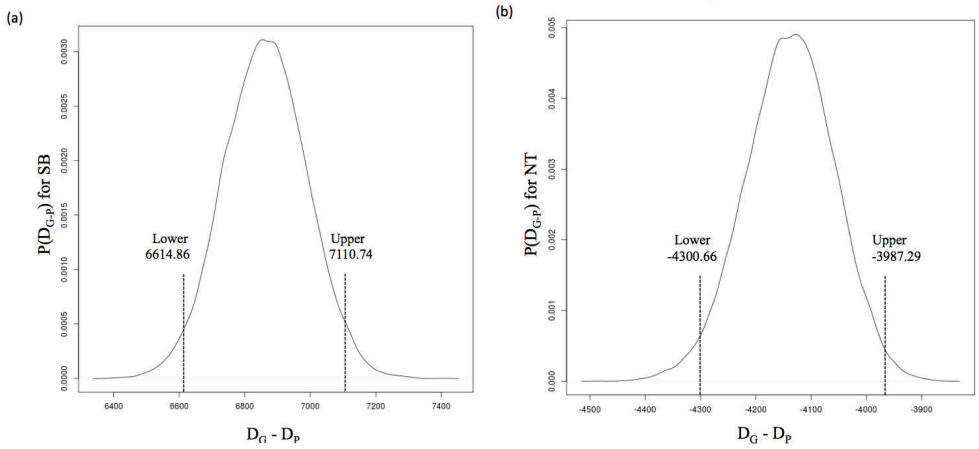


Figure 1 - Distribution of deviance difference plots between Gaussian and Poisson models. Plot (a) shows the distribution of deviance difference $P(D_{G-P})$ fitted to stillborn (SB) and (b) to number of teats (NT). D_G and D_P are the estimated Gaussian and Poisson models deviance, respectively. Positive deviance values (y axis) means that the Poisson distribution was the best fit.

Based on these results, the remainder of the analyses presented for SB uses the Poisson distribution, whilst the Gaussian distribution is used for NT. Using these distributions, we identified 18 SNPs related to SB and 65 to NT (listed in Additional file 1 and 2: Tables S1 and Table S2). Significant SNPs were accessed by 95% HPD intervals and posterior probability (PPN0). Under the HPD approach, if the value zero was not included in the interval for the marker effect, this marker was declared as significant. Additionally, if the PPN0 values were higher than 0.95 this significance was also reported.

From the significant SNPs we identified one QTL block including 4 SNPs on chromosome (chr) 1 for SB and nine QTL blocks for NT with three on chr 7 (4, 7 and 6 significant SNPs in each block, respectively), five at chr 8 (6, 4, 5, 4 and 5 significant SNPs in each block, respectively) and one at chr 12 including 4 SNPs (Additional file 3 and 4: Figure S1 and S2). 14 significant SNPs for SB and 20 for NT were unlinked to other significant SNPs. Based on these linkage/QTL blocks plus single markers location, we found 18 and 57 genes overlapping the significant SNP regions, respectively for SB and NT (Tables 1 and 2).

Table 1 - Significant SNPs for stillborn and associated genes. The table shows the significant SNPs, their positions in base pairs (bp) at swine chromosome (Chr), the QTL Block, associated genes (genes located in the Block or in an interval of 32,5 Kbp around each Block or SNP) followed by their distance in base pair of a single marker or in relation to the first or last SNP of the block.

SNP	Chr	Position (bp)	Block	Gene	Distance (bp)*
BGIS0003207	1	183948686	-	FEM1B	14336
				PIAS1	Inside
				ITGA11	6384
MARC0007670	1	193689396	-	-	-
M1GA0001259	1	204647860	Block 1	SAMD4A	6863
ALGA0007251	1	204871282		LOC102168145, WDHD1, SOCS4, MAPK1IP1L, LGALS3 and DLGAP5	Inside
MARC0056056	1	204934912			
H3GA0003422	1	205020361			
ASGA0010665	2	87274373	-	F2R	5412
				LOC102165085/IQGAP2-like	30878
ALGA0014165	2	87624560	-	LOC100520366	Inside
MARC0000488	2	87728463	-	-	-
ALGA0014249	2	88859718	-	-	-
ALGA0018674	3	47957752	-	NPHP1	Inside
				LOC102166895	9013
ASGA0028841	6	82315974	-	PTP4A2	19505
ASGA0070042	15	90388740	-	-	-
ALGA0086491	15	111088143	-	LOC100738946/HECW2	Inside
ASGA0070213	15	111116466	-	LOC100738946/HECW2	Inside
ASGA0072736	16	26077922	-	-	-
H3GA0049633	17	61860123	-	CYP24A1	31229
ASGA0077857	17	61889054	-	CYP24A1	2298

*Gene location with respect to the block, or the location of a single marker to the gene

Table 2 - Significant SNPs for number of teats and associated genes. The table shows the significant SNPs, their positions in base pairs (bp) at swine chromosome (Chr), the QTL Block, associated genes (genes located in the Block or in an interval of 32,5 Kbp around each Block or SNP) followed by their distance in base pair of a single marker or in relation to the first or last SNP of the block

SNP	Chr	Position (bp)	Block	Gene	Distance (bp)*
ALGA0004864	1	99713078	-	-	-
ALGA0012925	2	34084545	-	-	-
ALGA0012930	2	34177927	-	-	-
ALGA0013045	2	40181443	-	LOC102167107	20750
MARC0055904	2	40328584	-	SLC17A6	Inside
ASGA0032215	7	31600286	Block 1	KLHL31	978
H3GA0020592	7	31714979		LOC102166539, LOC102166618, GCLC, LOC102167236 and LOC102167364	Inside
MARC0010879	7	31869398			
MARC0098266	7	31945954		KHDRBS2	12332
ALGA0039995	7	32047280	Block 2	KHDRBS2	Inside
ALGA0040000	7	32134452			
ASGA0032254	7	32166462			
ASGA0032255	7	32192051			
MARC0043689	7	32252888			
INRA0024655	7	32313430			
ASGA0032266	7	32543114		-	-
ALGA0040040	7	32915748	-	PRIM2	Inside
ASGA0034811	7	91149363	-	-	-
H3GA0022644	7	102901720	-	PTGR2	27013
MARC0038565	7	103495170	-	VRTN	28094
				SYNDIG1L	5675
MARC0048752	7	103789642	Block 3	AREL1	5892
M1GA0010654	7	103796933		FCF1, YLPM1, PROX2, DLST and RPS6KL1	Inside
ALGA0043962	7	103816521			
H3GA0022664	7	103910821			
ASGA0035527	7	103933199			
DIAS0001088	7	103960033			
M1GA0010658	7	103999954	-	LOC102167367	26610
				PGF	Inside
ASGA0035536	7	104108293	-	MLH3	16923
				LOC102167860	5573
				ACYP1	Inside
				ZC2HC1C	1044
				NEK9	11806
ALGA0122954	7	104598913	-	JDP2	Inside
				LOC100624918/FLVCR2	8951
ASGA0035556	7	105224235	-	TGFB3	14323
				IFT43	Inside

Table 1 (Continue).

MARC0093074	8	50223543	Block 1	LOC102166479, C8H4orf45 and RAPGEF2	Inside
H3GA0024861	8	50329649			
H3GA0024862	8	50359681			
H3GA0024868	8	50479231			
H3GA0052920	8	50503562			
ASGA0038804	8	50537893			
DRGA0008588	8	51580681	-	-	-
MARC0077695	8	53929233	-	-	-
ALGA0047895	8	55647917	Block 2	LOC102159670	Inside
H3GA0024880	8	55670008			
H3GA0024879	8	55749069			
ALGA0047896	8	56064449			
ASGA0038818	8	56175366	-	-	-
H3GA0024884	8	56642918	Block 3	-	-
ASGA0038820	8	56673496			
H3GA0024882	8	56695265			
ASGA0038822	8	56764454			
ALGA0047901	8	56805057			
MARC0013221	8	57262741	-	LOC100519853/ZNF320	Inside
				LOC102160850	692
INRA0029832	8	58466955	Block 4	LOC102165777, NMU, LOC102166140, PDCL2, LOC102166866, CLOCK and LOC100523623/TMEM165	Inside
ALGA0047932	8	58557889			
ALGA0047933	8	58625849			
M1GA0011944	8	58807576		SRD5A3	13606
MARC0000554	8	67026060	-	LOC100737487/EPHA5-like	Inside
ASGA0085207	8	69065977	Block 5	-	-
MARC0020237	8	69070421			
MARC0095739	8	69146481			
ALGA0103392	8	69146919			
ALGA0102491	8	69215722			
ALGA0066725	12	50281949	Block 1	METTL16, PAFAH1B1, LOC102165360, CLUH and LOC102165602/CCDC92	Inside
ASGA0054883	12	50340265			
MARC0027202	12	50489072			
ALGA0066740	12	50578018			
DIAS0001557	12	55962023	-	PFAS	11951
				RANGRF	Inside
				SLC25A35	163
				ARHGEF15	8954
				ODF4	31445
ASGA0062949	14	43688399	-	TRPV4	4852
				FAM222A	4716

* Gene location with respect to the block, or the location of a single marker to the gene

Gene -TF Networks

Information about biological process, cellular component and molecular function of all identified genes, based on human gene annotation as it is annotated more comprehensively, are described in Additional file 5: Table S3 and Additional file 6: Table S4. Using the TFM-explorer, the regulatory sequence analyses were performed for each set of genes, identifying transcription factors (TF) strongly related (p-value<0.0001) with them (Additional file 7: Table S5 and Additional file 8: Table S6). The most representative TF (*FOXA1, FOXD1, NF-kappaB* and *KLF4* for SB and *ARNT, ELF5, RXRA::VDR* and *SOX9* for NT), based on their biological process and literature review, (e.g. TFs involved with kidney disease and diabetes on SB trait and related with mammary gland tissue and vertebrae composition and spinal cord expression on NT group) from each set of genes were chosen to generate a gene-TF network (Figures 2 and 3). According these networks we were able to point out the most probable candidate genes for SB (*CYP24A1, DLGAP5, F2R, IQGAP2, LGALS3, MAPK1IP1L, NPHP1, PTP4A2, PIAS1* and *WDHD1*) and NT (*PRIM2, AREL1, CLOCK, NEK9, NMU, SYNDIG1L, TGFB3, TMEM165-like, VRTN* and *YLPM1*).

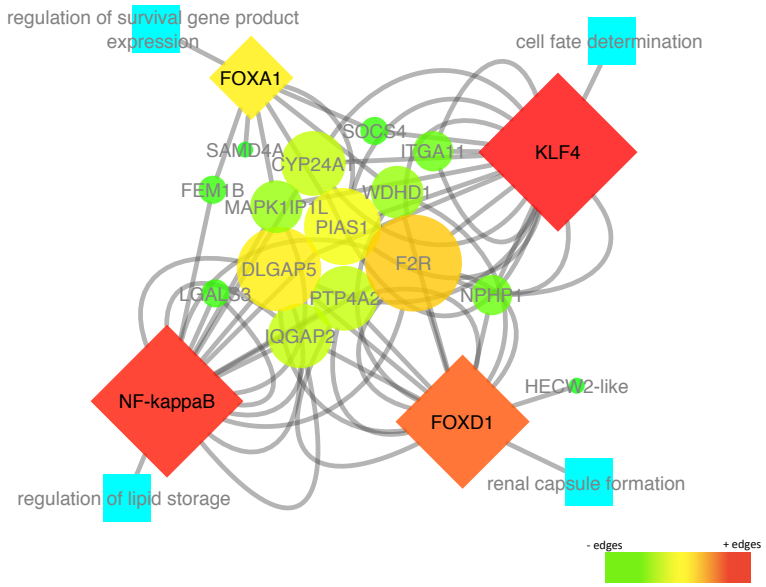


Figure 2 - Stillborn gene-Transcription Factor (TF) network. Four TF associated with stillborn genes: *NF-kappaB*, *FOXA1*, *KLF4* and *FOXD1* (diamond nodes) with *in silico* validated targets (circle nodes). Their color scale and size corresponds to network analyses (Cytoscape) score where red and bigger nodes represent higher edges density while green and smallest represent lower edges density. Blue nodes are the TF related biological process.

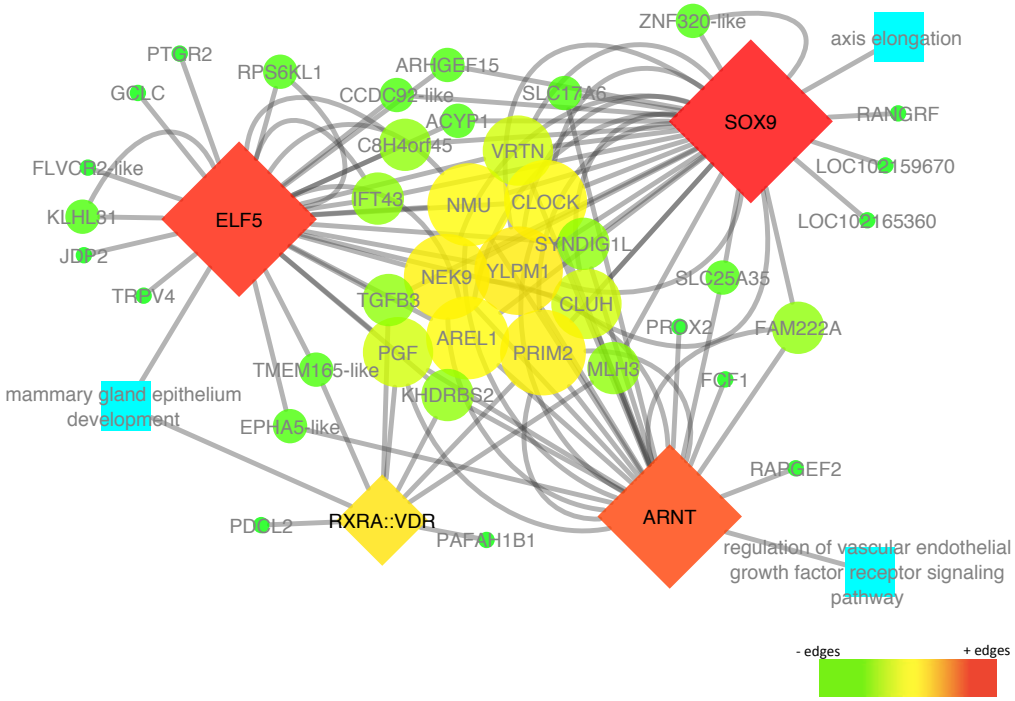


Figure 3 - Number of teats gene-Transcription Factor (TF) network. Four TF associated with number of teats genes: *SOX9*, *ELF5*, *RXRA::VDR* and *ARNT* (diamond nodes) with *in silico* validated target genes (circle nodes). Their color scale and size corresponds to network analyses (Cytoscape) score where red and bigger nodes represent higher edges density while green and smallest represent lower edges density. Blue nodes are the TF related biological process.

Discussion

The posterior distribution of the deviance difference ($P(D_{G-P})$) for each trait was assessed and used to infer about the goodness of fit of the tested models. The Poisson model presented the best fit for SB, while the Gaussian model was the best for the NT on this study (the density of SB and NT data are provided considering the observed, Poisson and Gaussian curve at Additional file 9 and 10: Figure S3 and S4 respectively). Most studies do not consider this particular distribution of SB [45], even though the use of different discrete models, such as Poisson and binomial, can lead to more appropriate quantification of the genetic influences on this trait. Working under a hierarchical Bayesian approach, Varona and Sorensen [46] also show the importance to propose and compare discrete models to be fitted to stillborn data under a genetic and animal

breeding approach. Thus, using genomic information (i.e. under GWAS or GWS approaches), the predictions of GEBVs and marker effects estimation exploiting counting models increased our analyses accuracy.

Even though NT it is also characterized as a counting discrete variable, the Gaussian distribution fitted the behavior of this trait better than the Poisson distribution. A possible reason is that the Poisson distribution is asymmetric and right skewed, and the symmetry of the Gaussian distribution was more consistent with the observed distribution of the NT sample data. Another possible explanation is that the Poisson distribution assumes the mean equal to its variance, a condition that may not have been met when working with NT. Besides, the mean for NT is much higher (15.3) than for SB (1.2), and the Gaussian distribution is a reasonable approximation for the Poisson when mean is greater than 10. In future studies, it would be interesting to consider other distributions used for discrete random variables for which such constraints are not required, e.g. the negative binomial and generalized Poisson distributions.

The distribution analyses showing the best fit of the Poisson distribution for SB and Gaussian for NT allowed us to perform the GWAS with more confidence. Thus, the GWAS for the SB trait indicated 18 significant SNPs that were distributed on chromosomes 1, 2, 3, 6, 15, 16 and 17, with one QTL block on chromosome 1 (Table 1). Several QTL for SB were previously reported near to the chromosome regions identified in this study for chromosome 1, 6, 16 and 17 [2]. For NT trait, the GWAS indicated 65 significant SNPs that were distributed in chromosomes 1, 2, 6, 9, 11 and 14, with QTL blocks identified on chromosome 7, 8 and 12 (Table 2). In different experimental populations, several QTLs for NT were previously reported in the same chromosomes regions identified in our study [47]-[51]. The identification of these markers associated with SB and NT traits in the GWAS and the confirmation of QTL identified in other studies gave us more confidence to evaluate the marker effects and

their biological function using post GWAS analyses as the gene-TF networks, and so, allowing to make a link between the traits, QTL and the overlapping genes.

Gene -TF Networks

Based on single markers and QTLs blocks we identified 18 and 57 genes, respectively, for SB and NT. From these genes we collected information about their GO terms and pathways e.g., renal system and reproductive biological process for SB genes and glandular epithelial cell and mammary gland development biological process for NT genes. The set of genes was used to explore the gene's promoter regions for enriched transcription factors (TF) in which the most representative, based on their biological process and literature review, for each trait were used to generate gene-TF networks highlighting the most probable candidate genes in each scenario.

Stillborn

Three of the four TF found on SB gene-TF network (*NF-kappaB, FOXA1* and *FOXD1*) are cited to be involved in (human) kidney disease and diabetes [52]-[54]. These mother diseases have been reported to be associated in elevated risk of stillbirth in humans [3]-[5]. The fourth TF (*KLF4*) is related with vascular and cardiac disease [55]. Heart disease is one of the major causes of mortality and morbidity in the human perinatal period [5]. Among pregnant diabetic women, fetal hypoxia and cardiac dysfunction secondary to poor glycaemic control are suggested important pathogenic factors in stillbirths [4].

With these TFs we could construct a network pointing out new candidate genes for SB (*F2R, IQGAP2, PTP4A2, WDHD1, PIAS1, DLGAP5, MAPK1IP1L, SOCS4, NPHP1* and *CYP24A1*). The *DLGAP5* gene was one of the most connected (highlighted) in the gene-TF network. *DLGAP5* has been cited to be a strong predictor of adrenocortical tumors [56]; and adrenocortical functions have been linked to renal diseases [57] that

are associated with SB occurrence. Also, a well-connected gene was the *F2R* that has a role in congenital heart disease [58]. In the network these two genes are connected to each other through the *NF-kappaB, KLF4* and *FOXD1* TF that are related with kidney and heart disease, hereby linking them with the stillborn trait.

As mentioned above in humans, a diabetic pregnancy influences the occurrence of stillbirths. In our SB network, one of the well-connected genes was the *CYP24A1* gene, which has been found to have a significantly higher expression in placental tissue from woman with gestation diabetes mellitus [59]. Other genes in the SB network related with diabetes are *PTP4A2, IQGAP2* and *SOCS4* [60]-[62]. Another strong candidate gene identified was nephronophthisis 1 (*NPHP1*), which is associated with juvenile nephronophthisis and Joubert Syndrome (JS) [63]. One of the key symptoms of JS is the breathing abnormality during neonatal period [64] that also could be linked with the occurrence of stillborn piglets. Using this gene-TF network we could confirm genes linked with SB trait not only through their position at a QTL but also through a known biological role.

Number of teats

TF *SOX9 and ARNT* from the NT gene-TF network are mainly involved with vertebrae composition and spinal cord expression. For example, the *SOX9* has been cited to be involved with campomelic dysplasia [65]; a syndrome which, among others, is characterized by vertebrae malformation and lower number of ribs [66]. *ARNT*, an aryl hydrocarbon nuclear regulatory factor, has been found to be expressed in the spinal cord during mouse development [67] and *ELF5* and *VDR* are cited to be involved with the mammary gland tissue [68], [69] being, in the Gene Ontology analyses, related to mammary gland development biological process; as illustrated in the gene-TF network (figure 3).

With these TFs we could construct a network pointing out potential new candidate genes for NT (*PRIM2, AREL1, YLPM1, NEK9, NMU, CLOCK, SYNDIG1L, TMEM165-like, TGFB3* and *VRTN*). Of these genes, well connected in the in the network *YLPM1* is a gene playing a role in the reduction of telomerase activity during differentiation of embryonic stem cells [70]. Also well connected, *PRIM2* is a DNA primase (large subunit) where a significant SNP associated to body length in a GWAS for Large White × Minzhu Pig Population has been found [71]. Linked with this group of genes we observe genes related with number of vertebrae, such as *PROX2, VRTN* and *SYNDIG1L* [72], [51], which in pigs may be correlated with number of teats [8], [31]. The chromosome regions of these three genes have been well explored due their observed significance for number of teats [31], [51] according our network. In addition, other genes which are well linked in the gene-TF network such as *NMU*, has been cited to be related to bone formation [73] and *TGFB3* has a significant association with the ossification of the posterior longitudinal ligament of the spine in humans [74]. Connected with two TF associated with mammary gland epithelium development GO term, *TMEM165-like* has been linked with developing, lactating and involuting mammary gland [75]. These genes, as others that have been better studied at molecular level (e.g. *AREL1, NEK9* and *CLOCK*), are highlighted in our network as the most probable candidate genes for NT.

Conclusions

We showed a better fit of the Poisson distribution for stillborn trait and superiority of the Gaussian for number of teats. We recommend that these distributions together with other discrete distributions should be tested in order to better evaluate count traits in GWAS. Based on these, it was also possible to observe associations between significant SNPs for these traits and genes mapped to them. The present study also provides

information about these genes increasing our understanding of the molecular mechanisms underlying then.

In addition, we predicted gene interactions that were coherent with known newborn survival traits and breast biology in mammals providing candidate genes for stillborn (e.g. *DLGAP5, PTP4A2, IQGAP2, SOCS4, CYP24A1, F2R* and *NPHP1*) and number of teats (e.g. *YLPM1, PROX2, VRTN, SYNDIG1L, PRIM2, TMEM165-like, NMU* and *TGFB3*). Our results highlighted important TF that may have an important role on the traits studied (e.g. *NF-kappaB* and *KLF4* for SB and *SOX9* and *ELF5* for NT). Nevertheless, these are complex traits that are subject to the action of a large number of genes that are regulated by several transcription factors, many of them still to be identified.

Abbreviations

NT, number of teats; SB, number of stillborn piglets; SNP, Single Nucleotide Polymorphisms; GWAS, genome wide association studies; QTL, Quantitative Trait Loci; TF, transcription factor; TFBS, transcription factor binding sites; HPD, highest posterior density; JS, Joubert Syndrome.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LLV, FFS and SEFG planned the experiment. LLV, FFS, MSL and MK ran the analyses. LLV, FFS, MSL, OM, SEFG, JWMB, PSL, EFK, LV, MD and MK contributed to drafting the manuscript. PSL and EK contributed to the conception of the study and provided data to the writing of the paper. All authors read and approved the final version.

Acknowledgements

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), National Institute of Science and Technology – Animal Science, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/NUFFIC and CAPES/DGU) and Institute for Pig Genetics - IPG (TOPIGS).

References

1. Blasco A, Bidanel JP, HaleyCS: **Genetics and neonatal survival**. In: Varley MA, editors. *The neonatal pig: development and survival*. CAB Int, Wallingford, UK; 1995. p. 17-38.

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

3. Nevis IF, Reitsma A, Dominic A, McDonald S, Thabane L, Akl EA, et al.: **Pregnancy Outcomes in Women with Chronic Kidney Disease: A Systematic Review**. *Clin J Am Soc Nephrol* 2011, **6**(11):2587-2598.

4. Mathiesen ER, Ringholm L, Damm P: **Stillbirth in diabetic pregnancies**. *Best Pract Res Clin Obstet Gynaecol* 2011, **25**(1):105-111.

5. Lee K-sun, Khoshnood B, Chen L, Stephen NW, Cromie JW, Mittendorf RL: **Infant mortality from congenital malformations in the United States, 1970 –1997**.*Obstet Gynecol* 2001, **98**:620-627.

6. Hirooka H, de Koning DJ, Harlizius B, van Arendonk JA, Rattink AP, Groenen MA, et al.: **A whole-genome scan for quantitative trait loci affecting teat number in pigs**. *J Anim Sci* 2001, **79**(9):2320-2326.

7. Hens JR, Wysolmerski JJ: **Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland**. *Breast Cancer Res* 2005, **7**(5):220.

8. Ren DR, Ren J, Ruan GF, Guo YM, Wu LH, Yang GC, et al.: **Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc × Chinese Erhualian intercross resource population**. *Anim Genet* 2012, **43**(5):545-551.

9. Uimari P, Sironen A, Sevón-Aimonen ML: **Whole-genome SNP association analysis of reproduction traits in the Finnish Landrace pig breed**. *Genet Sel Evol* 2011, **43**:42.

10. Schneider JF, Rempel LA, Rohrer GA: **Genome-wide association study of swine farrowing traits. Part I: Genetic and genomic parameter estimates**. *J Anim Sci* 2012, **90**:3353-3359.

11. Perez-Enciso M, Tempelman RJ, Gianola D: **A comparison between linear and Poisson mixed models for litter size in Iberian Pigs**. *Livest Prod Sci* 1993, **35**:303.

12. Ayres DR, Pereira RJ, Boligon AA, Silva FF, Schenkel FS, Roso VM, et al.: **Linear and Poisson models for genetic evaluation of tick resistance in cross-bred Hereford x Nellore cattle**. *J Anim Breed Genet* 2013, **130**(6):417-424.

13. Cui Y, Kim D-Y, Zhu J: **On the Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci With Count Data**. *Genetics* 2006, **174**(4):2159-2172.

14. Silva KM, Knol EF, Merks JWM, Guimarães SEF, Bastiaansen JWM, van Arendonk JAM, et al.: **Meta-analysis of results from quantitative trait loci mapping studies on pig chromosome 4**. *Anim Genet* 2011, **42**:280-292

15. Hadfield JD, Nakagawa S: **General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters.** *J Evol Biol* 2010, **23**:494-508.

16. Van Raden PM: **Efficient methods to compute genomic predictions**. *J Dairy Sci* 2008, **91**(11):4414-4423.

17. Stranden I, Garrick DJ: **Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit**. *J Dairy Sci* 2009, **92**:2971–2975.

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

19. Li Z, Gopal V, Li X, Davis JM, Casella G: **Simultaneous SNP identification in association studies with missing data**. *Ann App Stat* 2012, **6**(2):432-456.

20. Ramírez O, Quintanilla R, Varona L, Gallardo D, Díaz I, Pena RN, Amills M: **DECR1 and ME1 genotypes are associated with lipid composition traits in Duroc pigs**. *J Anim Breed Genet* 2014, **131**(1):46-52.

21. Cecchinato A, Ribeca C, Chessa S, Cipolat-Gotet C, Maretto F, Casellas J, Bittante G: **Candidate gene association analysis for milk yield, composition, urea nitrogen and somatic cell scores in Brown Swiss cows.** *Animal* 2014, **7**:1-9.

22. Mirina A, Atzmon G, Ye K, Bergman A: **Gene size matters**. *PLoS One* 2012, **7**(11):e49093.

23. Yu TP, Tuggle CK, Schmitz CB, Rothschild MF: **Association of PIT1 polymorphisms with growth and carcass traits in pigs**. *J Anim Sci* 1995, **73**(5):1282-1288.

24. Chen J, Yang XJ, Xia D, Wegner J, Jiang Z, Zhao RQ: **Sterol regulatory element binding transcription factor 1 expression and genetic polymorphism significantly affect intramuscular fat deposition in the longissimus muscle of Erhualian and Sutai pigs**. *J Anim Sci* 2008, **86**(1):57-63.

25. Fortes MR, Reverter-Gomez T, Hiriyur-Nagaraj S, Zhang Y, Jonsson N, Barris W, et al.:**A SNP-derived regulatory gene network underlying puberty in two tropical breeds of beef cattle**. *J Anim Sci* 2011, **89**:1669-1683.

26. Reverter A, Fortes MRS: **Building single nucleotide polymorphism-derived gene regulatory networks: Towards functional genomewide association studies**. *J Anim Sci* 2013, **91**:530-536.

27. Verardo LL, Silva FF, Varona L, Resende MDV, Bastiaansen JWM, Lopes PS, et al.: **Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs**. *J App Genet* 2015, **56**(1):123-32.

28. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al.: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology**. *PLoS One* 2009, **4**:e6524.

29. Oliphant A, Barker DL, Stuelpnagel JR, Chee MS: **BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping**. *Biotechniques* 2002, **32**(6):56-58.

30. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al.: **Analyses of pig genomes provide insight into porcine demography and evolution**. *Nature* 2012, **491**(7424):393-398.

31. Lopes MS, Bastiaansen JW, Harlizius B, Knol EF, Bovenhuis H: **A Genome-Wide Association Study Reveals Dominance Effects on Number of Teats in Pigs**. *PloS One* 2014, **9**(8):e105867.

32. Hadfield JD: **MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package**. *J Stat Softw* 2010, **33**(2):1-22.

33. Lorenzo-Bermejo J, Beckmann L, Chang-Claude J, Fischer C: **Using the posterior distribution of deviance to measure evidence of association for rare susceptibility variants**. *BMC Proc* 2011, **5**(9):S38.

34. Plummer M, Best N, Cowles K, Vines K: **CODA: Convergence diagnosis and output analysis for MCMC**. *R news* 2006, **6**(1):7-11.

35. Aguilar I, Misztal I, Legarra A, and Tsuruta S: **Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation**. *J Anim Breed Genet* 2011, **128**:422-428.

36. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**:263-265.

37. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al.: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**:2225-9.

38. Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models**. *Genetics* 1964, **49**:49-67.

39. Veroneze R, Bastiaansen JW, Knol EF, Guimarães SE, Silva FF, Harlizius B et al.: **Linkage disequilibrium patterns and persistence of phase in purebred and**

crossbred pig (Sus scrofa) populations. *BMC Genetics* 2014, **15**(1):126.

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

41. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, (32 Database):D91-4.

42. Touzet H, Varré JS: **Efficient and accurate P-value computation for Position Weight Matrices**. *Algorithms Mol Biol*. 2007, **2**(1510.1186):1748-7188.

43. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al.: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

44. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448-3449.

45. Canario L, Cantoni E, Le Bihan E, Caritez JC, Billon Y, Bidanel JP, et al.: **Between-breed variability of stillbirth and its relationship with sow and piglet characteristics**. *J Anim Sci* 2006, **84**:3185–3196.

46. Varona L, Sorensen D: **A genetic analysis of mortality in pigs**. *Genetics* 2010, **184**:277–284.

47. Guo Y-M, Lee GJ, Archibald AL, Haley CS: **Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan x Large White populations**. *Anim Genet* 2008, **39**(5):486-495

48. Lee SS, Chen Y, Moran C, Cepica S, Reiner G, Bartenschlager H, et al.: **Linkage and QTL mapping for Sus scrofa chromosome 2**. *J Anim Breed Genet* 2003, **120**(s1):11-19.

49. King AH, Jiang Z, Gibson JP, Haley CS, Archibald AL: **Mapping quantitative trait loci affecting female reproductive traits on porcine chromosome 8**. *Biol Reprod* 2003, **68**(6):2172-2179.

50. Sato S, Atsuji K, Saito N, Okitsu M, Komatsuda A, Mitsuhashi T, et al.: **Identification of quantitative trait loci affecting corpora lutea and number of teats in a Meishan× Duroc F2 resource population**. *J Animal Science* 2006, **84**(11):2895-2901.

51. Duijvesteijn N, Veltmaat JM, Knol EF, Harlizius B: **High-resolution association mapping of number of teats in pigs reveals regions controlling vertebral development**. *BMC genomics* 2014, **15**(1):542.

52. Mezzano S, Aros C, Droguett A, Burgos ME Ardiles L: **NF-KB activation and overexpression of regulated genes in human diabetic nephropathy**. *Nephrol Dial Transplant* 2004, **19**:2505-2512.

53. Behr R, Brestelli J, Fulmer JT, Miyawaki N, Kleyman TR, Kaestner KH: **Mild nephrogenic diabetes insipidus caused by Foxa1 deficiency**. *J Biol Chem* 2004,

279(40):41936-41941.

54. Levinson RS, Batourina E, Choi C, Vorontchikhina M, Kitajewski J, Mendelsohn CL: **Foxd1-dependent signals control cellularity in the renal capsule, a structure required for normal renal development**. *Development* 2005, **132**(3):529-539.

55. Yoshida T, Yamashita M, Horimai C, Hayashi M: **Deletion of Krüppel-like factor 4 in endothelial and hematopoietic cells enhances neointimal formation following vascular injury**. *J Am Heart Assoc* 2014, **3**(1):e000622.

56. Fragoso MCB, Almeida MQ, Mazzuco TL, Mariani BM, Brito LP, Gonçalves TC et al.: **Combined expression of BUB1B, DLGAP5, and PINK1 as predictors of poor outcome in adrenocortical tumors: validation in a Brazilian cohort of adult and pediatric patients**. *Eur J Endocrinol* 2012, **166**(1):61-67.

57. Arregger AL, Cardoso EM, Zucchini A, Aguirre EC, Elbert A, Contreras LN: **Adrenocortical function in hypotensive patients with end stage renal disease**. *Steroids* 2014, **84**:57-63.

58. Gigante B, Bellis A, Visconti R, Marino M, Morisco C, Trimarco V, et al.: **Retrospective analysis of coagulation factor II receptor (F2R) sequence variation and coronary heart disease in hypertensive patients**. *Arterioscler Thromb Vasc Biol* 2007, **27**(5):1213-1219.

59. Cho GJ, Hong SC, Oh MJ, Kim HJ: **Vitamin D deficiency in gestational diabetes mellitus and the role of the placenta**. *Am J Obstet Gynecol* 2013, **209**(6):560.e1-8.

60. Kakoola DN, Curcio-Brint A, Lenchik NI, Gerling IC: **Molecular pathway alterations in CD4 T-cells of nonobese diabetic (NOD) mice in the preinsulitis phase of autoimmune diabetes**. *Results Immunol* 2014, **4**:30-45.

61. Woroniecka KI, Park ASD, Mohtat D, Thomas DB, Pullman JM, Susztak K: **Transcriptome analysis of human diabetic kidney disease**. *Diabetes* 2011, **60**(9):2354-2369.

62. Feng X, Tang H, Leng J, Jiang Q: **Suppressors of cytokine signaling (SOCS) and type 2 diabetes**. *Mol Biol Rep* 2014, **41**(4):2265-2274.

63. Parisi MA, Bennett CL, Eckert ML, Dobyns WB, Gleeson JG, Shaw DW, et al.: **The NPHP1 gene deletion associated with juvenile nephronophthisis is present in a subset of individuals with Joubert syndrome**. *Am J Hum Genet* 2004, **75**(1):82-91.

64. Saravia JM, Baraister M: **Joubert syndrome: A review**. *Am J Med Genet* 1992, **43**:726-731.

65. Wunderle VM, Critcher R, Hastie N, Goodfellow PN, Schedl A: **Deletion of long-range regulatory elements upstream of SOX9 causes campomelic dysplasia**. *Proc Nat Acad Sci* 1998, **95**(18):10649-10654.

66. Mansour S, Hall CM, Pembrey ME, Young ID: **A clinical and genetic study of campomelic dysplasia**. *J Med Genet* 1995, **32**:415-420.

67. Jain S, Maltepe E, Lu MM, Simon C, Bradfield CA: **Expression of ARNT, ARNT2, HIF1α, HIF2 α and Ah receptor mRNAs in the developing mouse**. *Mech Dev* 1998, **73**(1):117-123.

68. Choi YS, Chakrabarti R, Escamilla-Hernandez R, Sinha S: **Elf5 conditional knockout mice reveal its role as a master regulator in mammary alveolar development: failure of Stat5 activation and functional differentiation in the absence of Elf5**. *Develop Biol* 2009, **329**(2):227-241.

69. Welsh J, Wietzke JA, Zinser GM, Smyczek S, Romu S, Tribble E, et al.: **Impact of the Vitamin D_3 receptor on growth-regulatory pathways in mammary gland and breast cancer**. *J Steroid Biochem Mol Biol* 2002, **83**(1):85-92.

70. Blalock WL, Piazzi M, Bavelloni A, Raffini M, Faenza I, D'Angelo A, et al.: **Identification of the PKR nuclear interactome reveals roles in ribosome biogenesis, mRNA processing and cell division**. *J Cel Physiol* 2014, **229**(8):1047-1060.

71. Wang L, Zhang L, Yan H, Liu X, Li N, Liang J, et al.: **Genome-Wide Association Studies Identify the Loci for 5 Exterior Traits in a Large White× Minzhu Pig Population**. *PloS One* 2014, **9**(8):e103766.

72. Pistocchi A, Bartesaghi S, Cotelli F, Del Giacco L: **Identification and expression pattern of zebrafish prox2 during embryonic development**. *Develop Dyn* 2008, **237**(12):3916-3920.

73. Sato S, Hanada R, Kimura A, Abe T, Matsumoto T, Iwasaki M, et al.: **Central control of bone remodeling by neuromedin U.** *Nat Med* 2007, **13**(10):1234-1240.

74. Horikoshi T, Maeda K, Kawaguchi Y, Chiba K, Mori K, Koshizuka Y, et al.: **A large-scale genetic association study of ossification of the posterior longitudinal ligament of the spine**. *Hum Genet* 2006, **119**(6):611-616.

75. Reinhardt TA, Lippolis JD, Sacco RE: **The Ca^{2+}/H^{+} antiporter TMEM165 expression, localization in the developing, lactating and involuting mammary gland parallels the secretory pathway Ca^{2+} ATPase (SPCA1)**. *Biochem Biophys Res Commun* 2014, **445**(2):417-421.

Supplementary data

Additional file 1: Table S1 Significant SNPs, position and chromosome (chr) location on *S. scrofa* reference genome (10.2), posterior mean, posterior probability under H_0 (PPN0) and 95% HPD (Highest Posterior Density) interval limits for stillborn trait.

					95% HPD interval	
SNP	**chr**	**Position (bp)**	**mean**	**PPN0**	**Lower**	**Upper**
BGIS0003207	1	183948686	0.00100	0.96085	0.00003	0.002009
MARC0007670	1	193689396	-0.00073	0.95015	-0.00151	-0.000005
M1GA0001259	1	204647860	0.00088	0.96699	0.00003	0.001797
ALGA0007251	1	204871282	0.00089	0.97354	0.00004	0.001804
MARC0056056	1	204934912	0.00089	0.97856	0.00005	0.001796
H3GA0003422	1	205020361	0.00086	0.95455	0.00001	0.001769
ASGA0010665	2	87274373	0.00087	0.98175	0.00007	0.001716
ALGA0014165	2	87624560	0.00091	0.98798	0.00009	0.001783
MARC0000488	2	87728463	0.00094	0.99014	0.00012	0.001805
ALGA0014249	2	88859718	0.00089	0.98213	0.00007	0.001739
ALGA0018674	3	47957752	-0.00077	0.95105	-0.00159	-0.000002
ASGA0028841	6	82315974	-0.00104	0.95389	-0.00213	-0.000011
ASGA0070042	15	90388740	-0.00097	0.96842	-0.00196	-0.000045
ALGA0086491	15	111088143	0.00082	0.95388	0.00001	0.001654
ASGA0070213	15	111116466	0.00082	0.95455	0.00001	0.001653
ASGA0072736	16	26077922	-0.00102	0.96587	-0.00208	-0.000034
H3GA0049633	17	61860123	-0.00119	0.96875	-0.00240	-0.000047
ASGA0077857	17	61889054	-0.00115	0.95241	-0.00237	-0.000003

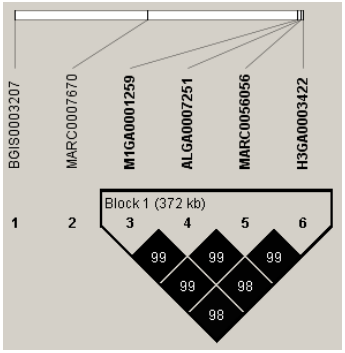
Additional file 2: Table S2 Significant SNPs, position and chromosome (chr) location on *S. scrofa* reference genome (10.2), posterior mean, posterior probability under H_0 (PPN0) and 95% HPD (Highest Posterior Density) interval limits for number of teats**.**

					95% HPD interval	
SNP	**chr**	**Position (bp)**	**mean**	**PPN0**	**Lower**	**Upper**
ALGA0004864	1	99713078	0.23669	0.9996	0.08023	0.39306
ALGA0012925	2	34084545	-0.02302	0.99065	-0.04290	-0.00307
ALGA0012930	2	34177927	0.02356	0.99500	0.00364	0.04337
ALGA0013045	2	40181443	0.00120	0.95014	0.00001	0.00245
MARC0055904	2	40328584	0.00121	0.95235	0.00001	0.00244
ASGA0032215	7	31600286	-0.00116	0.95403	-0.00228	-0.00009
H3GA0020592	7	31714979	0.00115	0.98034	0.00007	0.00229
MARC0010879	7	31869398	0.00121	0.99117	0.00013	0.00234
MARC0098266	7	31945954	-0.00108	0.95384	-0.00213	-0.00005
ALGA0039995	7	32047280	-0.00114	0.95391	-0.00222	-0.00010
ALGA0040000	7	32134452	-0.00121	0.95433	-0.00230	-0.00015
ASGA0032254	7	32166462	-0.00115	0.95399	-0.00223	-0.00009
ASGA0032255	7	32192051	-0.00112	0.95389	-0.00217	-0.00009
MARC0043689	7	32252888	-0.00124	0.95504	-0.00244	-0.00009
INRA0024655	7	32313430	-0.00119	0.95404	-0.00227	-0.00015
ASGA0032266	7	32543114	-0.00105	0.95383	-0.00213	-0.00001
ALGA0040040	7	32915748	-0.00112	0.95392	-0.00222	-0.00004
ASGA0034811	7	91149363	0.00126	0.97123	0.00005	0.00251
H3GA0022644	7	102901720	0.00246	0.99415	0.00051	0.00453
MARC0038565	7	103495170	-0.00239	0.99001	-0.00454	-0.00034
MARC0048752	7	103789642	0.00186	0.95996	0.00002	0.00377
M1GA0010654	7	103796933	0.00193	0.98797	0.00009	0.00384
ALGA0043962	7	103816521	0.00191	0.98467	0.00008	0.00382
H3GA0022664	7	103910821	0.00191	0.97346	0.00005	0.00383
ASGA0035527	7	103933199	0.00188	0.97054	0.00004	0.00379
DIAS0001088	7	103960033	0.00191	0.98677	0.00008	0.00383
M1GA0010658	7	103999954	0.00211	0.99411	0.00049	0.00383
ASGA0035536	7	104108293	0.00203	0.99222	0.00026	0.00390
ALGA0122954	7	104598913	0.00208	0.99421	0.00050	0.00376
ASGA0035556	7	105224235	0.00213	0.98544	0.00008	0.00431
MARC0093074	8	50223543	-0.00126	0.95612	-0.00251	-0.00007
H3GA0024861	8	50329649	-0.00130	0.95677	-0.00255	-0.00010
H3GA0024862	8	50359681	-0.00126	0.95616	-0.00251	-0.00006
H3GA0024868	8	50479231	-0.00129	0.95779	-0.00258	-0.00006
H3GA0052920	8	50503562	-0.00126	0.95622	-0.00251	-0.00007
ASGA0038804	8	50537893	-0.00128	0.95634	-0.00253	-0.00008
DRGA0008588	8	51580681	-0.00136	0.95714	-0.00260	-0.00017
MARC0077695	8	53929233	-0.00135	0.95765	-0.00259	-0.00017
ALGA0047895	8	55647917	-0.00105	0.95209	-0.00208	-0.00002
H3GA0024880	8	55670008	-0.00109	0.95384	-0.00212	-0.00006
H3GA0024879	8	55749069	-0.00112	0.95391	-0.00223	-0.00004
ALGA0047896	8	56064449	-0.00105	0.95376	-0.00208	-0.00002
ASGA0038818	8	56175366	-0.00105	0.95361	-0.00208	-0.00002
H3GA0024884	8	56642918	-0.00105	0.95371	-0.00210	-0.00001
ASGA0038820	8	56673496	-0.00107	0.95374	-0.00211	-0.00004
H3GA0024882	8	56695265	0.00109	0.97938	0.00006	0.00213
ASGA0038822	8	56764454	-0.00104	0.95376	-0.00208	-0.00001
ALGA0047901	8	56805057	-0.00109	0.95386	-0.00212	-0.00006
MARC0013221	8	57262741	-0.00105	0.95372	-0.00208	-0.00002
INRA0029832	8	58466955	-0.00105	0.95369	-0.00208	-0.00002
ALGA0047932	8	58557889	-0.00105	0.95371	-0.00208	-0.00002
ALGA0047933	8	58625849	0.00109	0.97808	0.00006	0.00213
M1GA0011944	8	58807576	0.00109	0.97898	0.00006	0.00213
MARC0000554	8	67026060	0.00098	0.97323	0.00005	0.00192
ASGA0085207	8	69065977	0.00099	0.96997	0.00004	0.00195
MARC0020237	8	69070421	0.00098	0.97004	0.00004	0.00193
MARC0095739	8	69146481	0.00096	0.95875	0.00002	0.00191

Table S2 Continue

ALGA0103392	8	69146919	0.00099	0.97042	0.00004	0.00195
ALGA0102491	8	69215722	0.00096	0.95903	0.00002	0.00191
ALGA0066725	12	50281949	0.00209	0.99322	0.00031	0.00395
ASGA0054883	12	50340265	0.00210	0.99324	0.00032	0.00396
MARC0027202	12	50489072	0.00214	0.99398	0.00038	0.00397
ALGA0066740	12	50578018	0.00185	0.98066	0.00007	0.00372
DIAS0001557	12	55962023	0.00188	0.97901	0.00006	0.00380
ASGA0062949	14	43688399	0.00143	0.98109	0.00007	0.00284

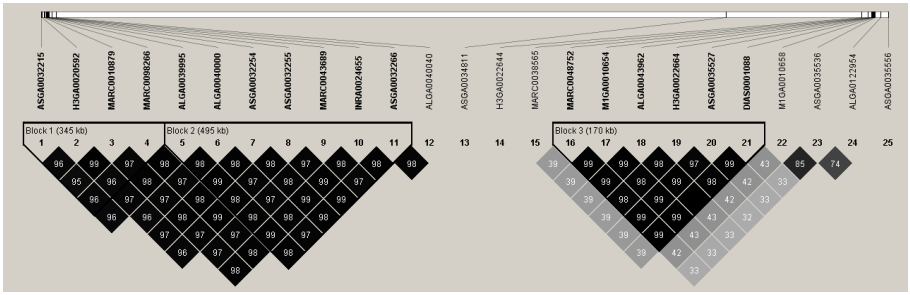
Chr 1



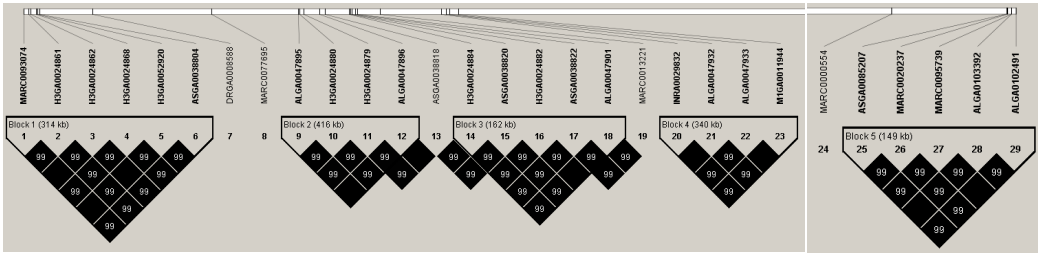
Additional file 3: Figure S1

QTL block present at chromosome 1 (Chr 1) that contain significant SNPs for stillborn. Solid lines mark the block.

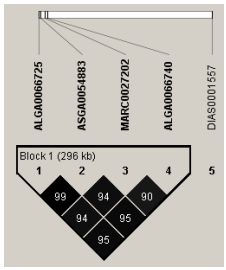
Chr 7



Chr 8



Chr 12



Additional file 4: Figure S2

QTL blocks present at chromosomes 7, 8 and 12 (Chr 7, Chr 8 and Chr 12 respectively) that contains significant SNPs for number of teats. Solid lines mark the blocks.

Additional file 5: Table S3 Pathway and Gene Ontology terms (GO) of genes associated with significant SNPs identified for stillborn.

Gene	Pathway	Molecular Function (GO)	Biological Process (GO)	Cellular Component (GO)
PIAS1	Cytokine Signaling in Immune system	transitional metal ion binding/protein-glutamic acid ligase activity	cellular protein catabolic process/reproductive process	nucleoplasm
CYP24A1	Metabolism of lipds and lipoproteins	transitional metal ion binding	fat-soluble vitamin metabolic process	mitochondrial part
PTP4A2	Signaling events mediated by PRL	phosphoprotein phosphatase activity	protein dephosphorylation	endosome
SAMD4A	-	translation regulator activity	regulation of translation	neuron projection
F2R	PI3K-Akt signaling pathway	protein complex binding	positive regulation of hydrolase activity/renal system process	intrinsic componet of plasma
DLGAP5	Aurora A signaling	phosphoprotein phosphatase activity	regulation of organelle organization	cytoskeletal part
WDHD1	-	chromatin binding	regulation of organelle organization	nucleoplasm
LGALS3	AGE/RAGE pathway	protein complex binding	extrinsic apoptotic signaling pathway	mitochondrial part
HECW2	-	protein-glutamic acid ligase activity	cellular protein catabolic process	-
IQGAP2	TNF-alpha/NF-kB Signaling pathway	enzyme activator activity	positive regulation of hydrolase activity	cytoskeletal part
ITGA11	PI3K-Akt signaling pathway	protein complex binding	muscle organ development	intrinsic componet of plasma
FEM1B	-	cytokine receptor binding	reproduction processs/extrinsic apoptotic signaling pathway	-
SOCS4	Type II diabetes mellitus	-	cellular protein catabolic process	-
NPHP1	-	structural molecule activity	reproduction processs	cell projection part
MAPK1IP1L	-	-	-	-
LOC102166895	-	-	-	-
LOC100520366	-	-	-	-
LOC102168145	-	-	-	-

Additional file 6: Table S4 Pathway and Gene Ontology terms (GO) of genes associated with significant SNPs identified for number of teats.

Gene	Pathway	Molecular Function (GO)	Biological Process (GO)	Cellular Component (GO)
TGFB3	TGF-beta Receptor Signaling Pathway	identical protein binding	protein phosphorylation/mammary gland development/positive regulation of signal transduction/regulation of multicelluar organismal development	extracellular space/neuron part
CLOCK	Metabolism of lipids and lipoproteins	nucleic acid binding transcription factor activity	glandular epithelial cell development/regulation of multicellular organismal development/reproduction	ribonucleoprotein complex
PGF	Rap1 signaling pathway	identical protein binding	regulation of multicellular organismal development	extracellular space
ACYP1	Glycolysis Gluconeogenesis	acylphosphatase activity	-	-
PRIM2	Cell cycle	helicase activity	DNA recombination	nucleoplasm
GCLC	Metabolism of amino acids and derivates	ATP binding	carbohydrate metabolic process	glutamate-cysteine ligase complex
SRD5A3	Metabolism of lipids and lipoproteins	oxidoreductase activity	carbohydrate metabolic process	endoplasmatic reticulum membrane
PAFAH1B1	Cell cycle	identical protein binding	positive regulation of signal transduction/regulation of multicellular organismal development/reproduction	microtubule organizing center/neuron part/ organelle envelope
RAPGEF2	Rap1 signaling pathway	guanyl-nucleotide exchange factor activity	protein phosphorylation/regulation of multicellular organismal development/positive regulation of signal transduction	neuron part
ARHGEF15	EPHA forward signaling	guanyl-nucleotide exchange factor activity	regulation of multicellular organismal development	neuron part
TMEM165	-	-	carbohydrate metabolic process	Golgi apparatus
TRPV4	Transmembrane transport of small molecules	ATP binding	regulation of multicellular organismal development	neuron part
MLH3	Meiosis	ATP binding	reproduction/DNA recombination	nuclear chromosome

Table S4. Continue

RANGRF	-	guanyl-nucleotide exchange factor activity	monovalent inorganic cation transport	nucleoplasm
EPHA5	EPHA forward signaling	ATP binding	positive regulation of signal transduction/protein phosphorilation/regulation of multicellular organismal development	neuron part
NMU	GPCR downstream signaling	neuropeptide receptor binding	feeding behavior	-
NEK9	Cell cycle	ATP binding	protein phosphorylation	
IFT43	-	-	cilium morphogenesis	microtubule organizing center
PFAS	Metabolism of nucleotides	ATP binding	nucleotide biosynthetic process	-
SLC17A6	Transmembrane transport of small molecules	active transmembrane transporter activity	monovalent inorganic cation transport	neuron part
DLST	Metabolism of amino acids and derivates	heat shock protein binding	generation of precursor metabolites and energy	transferase complex
SLC25A35	-		-	organelle envelop
KHDRBS2	-	SH3 domain binding	positive regulation of signal transduction	-
FLVCR2	-	tetrapyrrole binding	cofactor transport	-
ODF4	-	-	reproduction	ciliary part
FCF1	Ribosome biogenesis	poly(A) RNA binding	RNA processing	ribonucleoprotein complex
RPS6KL1	-	ATP binding	protein phosphorylation	ribonucleoprotein complex
METTL16	-	poly(A) RNA binding/methyltransferase activity	methylation	-
VRTN	-	trasposase activity	DNA recombination	-
YLPM1	-	poly(A) RNA binding/nucleic acid binding transcription factor activity	negative regulation of transcription from RNA polymerase II promoter	nucleoplasm
JDP2	Apoptosis signaling pathway	nucleic acid binding transcription factor activity	negative regulation of transcription from RNA polymerase II promoter	-
ZNF320	Gene expression	nucleic acid binding transcription factor activity	-	-
PTGR2	-	oxidoreductase activity	monocarboxylic acid metabolic process	-

Table S4. Continue

AREL1	-	ubiquitin-protein transferase activity	protein ubiquitination involved in ubiquitin-dependent protein catabolic process	-
CCDC92	-	-	-	microtubule organizing center
SYNDIG1L	-	-	response to biotic stimlus	Golgi apparatus
CLUH	Translation Factors	-	mitochondrion localization	neuron part
KLHL31	-	protein binding	regulation of transcription, DNA templated	-
PROX2	-	DNA binding	multicellular organismal development	nucleus
ZC2HC1C	-	metal ion binding	-	-
PDCL2	-	-	-	-
C8H4orf45	-	-	-	-
FAM222A	-	-	-	-
LOC102167107	-	-	-	-
LOC102166539	-	-	-	-
LOC102166618	-	-	-	-
LOC102167236	-	-	-	-
LOC102167364	-	-	-	-
LOC102167367	-	-	-	-
LOC102167860	-	-	-	-
LOC102166479	-	-	-	-
LOC102160850	-	-	-	-
LOC102165777	-	-	-	-
LOC102166140	-	-	-	-
LOC102165360	-	-	-	-
LOC102166866	-	-	-	-
LOC102159670	-	-	-	-

Additional file 7: Table S5 Transcription factors (TF) identified for stillborn genes, the location in base pairs (bp) of the window sequence presenting biding sites and the p-value of the window confidence obtained from TFM-explorer tool.

TF	Genes*	Location of the window (bp)		p-value
NKX3-1		929:26		0.000002
	FEM1B	152	145	
	PIAS1	904	897	
	ITGA11	905	898	
	ITGA11	676	669	
	ITGA11	531	524	
	ITGA11	512	505	
	ITGA11	500	493	
	ITGA11	492	485	
	ITGA11	449	442	
	ITGA11	238	231	
	ITGA11	42	35	
	WDHD1	527	520	
	SOCS4	899	892	
	LGALS3	929	922	
	DLGAP5	780	773	
	F2R	580	573	
	F2R	333	326	
	F2R	82	75	
	IQGAP2	481	474	
	IQGAP2	331	324	
	NPHP1	865	858	
	NPHP1	838	831	
	NPHP1	254	247	
	PTP4A2	867	860	
	PTP4A2	26	19	
	HECW2-like	919	912	
	HECW2-like	452	445	
En1		1443:1157		0.00002
	SAMD4A	1271	1260	
	WDHD1	1374	1363	
	WDHD1	1161	1150	
	SOCS4	1368	1357	
	MAPK1IP1L	1437	1426	
	DLGAP5	1443	1432	
	DLGAP5	1274	1263	
	DLGAP5	1177	1166	
	F2R	1272	1261	
	F2R	1175	1164	
	NPHP1	1336	1325	
	NPHP1	1196	1185	
	PTP4A2	1386	1375	
	HECW2-like	1157	1146	
	CYP24A1	1413	1402	

Table S5. Continue

NF-kappaB		1869:1714		0.00003
	FEM1B	1777	1767	
	PIAS1	1869	1859	
	MAPK1IP1L	1737	1727	
	DLGAP5	1867	1857	
	DLGAP5	1837	1827	
	DLGAP5	1714	1704	
	F2R	1858	1848	
	F2R	1747	1737	
	IQGAP2	1765	1755	
	IQGAP2	1715	1705	
	PTP4A2	1775	1765	
GABPA		2860:2709		0.00003
	PIAS1	2709	2698	
	ITGA11	2818	2807	
	SAMD4A	2732	2721	
	SOCS4	2792	2781	
	LGALS3	2806	2795	
	DLGAP5	2841	2830	
	NPHP1	2860	2849	
	NPHP1	2828	2817	
	HECW2-like	2803	2792	
	CYP24A1	2733	2722	
ELK1		2580:2002		0.00005
	FEM1B	2342	2332	
	ITGA11	2523	2513	
	ITGA11	2271	2261	
	SAMD4A	2232	2222	
	SAMD4A	2002	1992	
	WDHD1	2013	2003	
	DLGAP5	2303	2293	
	F2R	2580	2570	
	F2R	2049	2039	
	F2R	2024	2014	
	IQGAP2	2494	2484	
	IQGAP2	2132	2122	
	NPHP1	2317	2307	
	CYP24A1	2479	2469	
	CYP24A1	2059	2049	
SPIB		652:560		0.00007
	PIAS1	652	645	
	PIAS1	615	608	
	PIAS1	560	553	
	MAPK1IP1L	583	576	
	F2R	621	614	
	F2R	575	568	
	IQGAP2	615	608	
	NPHP1	631	624	
	HECW2-like	650	643	
	HECW2-like	593	586	

Table S5 Continue

FOXD1		382:15		0.00008
	WDHD1	259	251	
	WDHD1	241	233	
	WDHD1	230	222	
	DLGAP5	382	374	
	DLGAP5	349	341	
	F2R	353	345	
	F2R	286	278	
	F2R	15	7	
	IQGAP2	342	334	
	NPHP1	142	134	
	PTP4A2	249	241	
	PTP4A2	136	128	
	PTP4A2	77	69	
	PTP4A2	57	49	
	HECW2-like	326	318	
FOXA1		2996:2966		0.00009
	FEM1B	2996	2985	
	PIAS1	2993	2982	
	PIAS1	2966	2955	
	SAMD4A	2978	2967	
	WDHD1	2985	2974	
	SOCS4	2976	2965	
	MAPK1IP1L	2991	2980	
NFKB1		2507:2425		0.00005
	MAPK1IP1L	2492	2481	
	LGALS3	2447	2436	
	LGALS3	2428	2417	
	F2R	2439	2428	
	IQGAP2	2507	2496	
	IQGAP2	2462	2451	
	CYP24A1	2425	2414	
Klf4		2963:2736		0.0001
	PIAS1	2911	2901	
	PIAS1	2892	2882	
	PIAS1	2795	2785	
	ITGA11	2939	2929	
	ITGA11	2826	2816	
	ITGA11	2749	2739	
	SOCS4	2960	2950	
	MAPK1IP1L	2740	2730	
	DLGAP5	2746	2736	
	DLGAP5	2736	2726	
	F2R	2928	2918	
	F2R	2826	2816	
	F2R	2790	2780	
	NPHP1	2963	2953	
	NPHP1	2849	2839	
	CYP24A1	2895	2885	
	CYP24A1	2847	2837	
	CYP24A1	2773	2763	
	CYP24A1	2753	2743	

*Genes are repeated according to the number of transcription factor binding sites

Additional file 8: Table S6 Transcription factors (TF) identified for number of teats genes, the location in base pairs (bp) of the window sequence presenting binding sites and the p-value of the window confidence obtained from TFM-explorer tool.

TF	Genes*	Location of the window (bp)		p-value
HIF1A::ARNT		3122:2461		0.00000000008
	KHDRBS2	2556	2548	
	PRIM2	2873	2865	
	PRIM2	2762	2754	
	PRIM2	2569	2561	
	PRIM2	2503	2495	
	PRIM2	2495	2487	
	PTGR2	3086	3078	
	PTGR2	2675	2667	
	PROX2	2929	2921	
	RPS6KL1	3026	3018	
	RPS6KL1	2820	2812	
	RPS6KL1	2615	2607	
	MLH3	3122	3114	
	MLH3	3107	3099	
	MLH3	2461	2453	
	ACYP1	2872	2864	
	ACYP1	2502	2494	
	JDP2	2520	2512	
	FLVCR2-like	2762	2754	
	TGFB3	2893	2885	
	C8H4orf45	3087	3079	
	RAPGEF2	2513	2505	
	CLOCK	2844	2836	
	METTL16	2594	2586	
	LOC102165360	2889	2881	
	CLUH	2968	2960	
	CLUH	2677	2669	
	SLC25A35	2816	2808	
	SLC25A35	2787	2779	
	FAM222A	3070	3062	
	FAM222A	2924	2916	
Egr1		3127:2761		0.000000001
	KLHL31	3004	2993	
	GCLC	2848	2837	
	KHDRBS2	3082	3071	
	KHDRBS2	2959	2948	
	PRIM2	2801	2790	
	PRIM2	2788	2777	
	PRIM2	2761	2750	
	PTGR2	2770	2759	
	SYNDIG1L	3096	3085	
	SYNDIG1L	3047	3036	
	AREL1	3112	3101	
	YLPM1	3127	3116	

Table S6. Continue

	PROX2	3115	3104	
	RPS6KL1	3080	3069	
	RPS6KL1	2862	2851	
	PGF	3106	3095	
	MLH3	3031	3020	
	ACYP1	2977	2966	
	ACYP1	2927	2916	
	NEK9	2996	2985	
	NEK9	2964	2953	
	JDP2	2850	2839	
	JDP2	2836	2825	
	RAPGEF2	2815	2804	
	ZNF320-like	3097	3086	
	ZNF320-like	3073	3062	
	SRD5A3	2910	2899	
	EPHA5-like	2920	2909	
	METTL16	3113	3102	
	LOC102165360	2901	2890	
	LOC102165360	2791	2780	
	CCDC92-like	3110	3099	
	CCDC92-like	3048	3037	
	CCDC92-like	2964	2953	
	CCDC92-like	2923	2912	
	CCDC92-like	2866	2855	
	CCDC92-like	2835	2824	
	SLC25A35	2979	2968	
	TRPV4	2771	2760	
	FAM222A	3102	3091	
	FAM222A	3069	3058	
	FAM222A	3029	3018	
	FAM222A	2871	2860	
NFKB1		3130:2507		0.00000004
	SLC17A6	2753	2742	
	KLHL31	2530	2519	
	GCLC	2998	2987	
	GCLC	2578	2567	
	KHDRBS2	2912	2901	
	PRIM2	2986	2975	
	PRIM2	2716	2705	
	PRIM2	2617	2606	
	PRIM2	2554	2543	
	PTGR2	2753	2742	
	PTGR2	2551	2540	
	VRTN	3116	3105	
	VRTN	2740	2729	
	SYNDIG1L	3090	3079	
	SYNDIG1L	2674	2663	
	SYNDIG1L	2525	2514	
	PROX2	3098	3087	
	PROX2	2626	2615	

Table S6. Continue

	RPS6KL1	3077	3066	
	RPS6KL1	2808	2797	
	RPS6KL1	2748	2737	
	ACYP1	3130	3119	
	ACYP1	3083	3072	
	ACYP1	2739	2728	
	ZC2HC1C	3118	3107	
	ZC2HC1C	2909	2898	
	NEK9	3044	3033	
	NEK9	2982	2971	
	TGFB3	2975	2964	
	TGFB3	2828	2817	
	RAPGEF2	2806	2795	
	LOC102159670	2771	2760	
	ZNF320-like	3064	3053	
	ZNF320-like	3026	3015	
	ZNF320-like	3002	2991	
	NMU	3033	3022	
	NMU	2826	2815	
	CLOCK	2812	2801	
	TMEM165-like	2970	2959	
	TMEM165-like	2701	2690	
	PAFAH1B1	2741	2730	
	LOC102165360	2904	2893	
	LOC102165360	2657	2646	
	PFAS	3040	3029	
	PFAS	2741	2730	
	PFAS	2608	2597	
	RANGRF	3037	3026	
	RANGRF	2667	2656	
	SLC25A35	2808	2797	
	ARHGEF15	3004	2993	
	ARHGEF15	2703	2692	
	ODF4	2507	2496	
	TRPV4	2754	2743	
	TRPV4	2552	2541	
	FAM222A	2518	2507	
NF-kappaB			2558:2432	0.0000002
	KLHL31	2529	2519	
	GCLC	2451	2441	
	PRIM2	2554	2544	
	PRIM2	2435	2425	
	PTGR2	2551	2541	
	SYNDIG1L	2438	2428	
	PGF	2552	2542	
	MLH3	2470	2460	
	ACYP1	2498	2488	
	JDP2	2456	2446	
	FLVCR2-like	2432	2422	
	TMEM165-like	2434	2424	

Table S6. Continue

	CLUH	2472	2462	
	CCDC92-like	2470	2460	
	PFAS	2558	2548	
	RANGRF	2456	2446	
	RANGRF	2445	2435	
	SLC25A35	2435	2425	
	ODF4	2506	2496	
	TRPV4	2552	2542	
	FAM222A	2518	2508	
	FAM222A	2453	2443	
Hltf		1048:989		0.000002
	KHDRBS2	1033	1023	
	KHDRBS2	1017	1007	
	FCF1	1026	1016	
	RPS6KL1	1025	1015	
	PGF	1048	1038	
	JDP2	998	988	
	FLVCR2-like	1022	1012	
	FLVCR2-like	989	979	
	C8H4orf45	1042	1032	
	C8H4orf45	1028	1018	
	LOC102159670	1035	1025	
	LOC102159670	994	984	
	ZNF320-like	1031	1021	
	PDCL2	1011	1001	
	LOC102166866	1032	1022	
	PAFAH1B1	1013	1003	
	RANGRF	1034	1024	
Arnt		2980:2176		0.000004
	SLC17A6	2768	2762	
	KHDRBS2	2819	2813	
	PRIM2	2590	2584	
	PRIM2	2568	2562	
	PRIM2	2443	2437	
	VRTN	2236	2230	
	SYNDIG1L	2231	2225	
	AREL1	2250	2244	
	FCF1	2980	2974	
	YLPM1	2953	2947	
	PROX2	2928	2922	
	PGF	2478	2472	
	PGF	2356	2350	
	MLH3	2866	2860	
	NEK9	2760	2754	
	NEK9	2572	2566	
	NEK9	2402	2396	
	RAPGEF2	2512	2506	
	NMU	2571	2565	
	NMU	2183	2177	
	CLOCK	2665	2659	

Table S6. Continue

	CLOCK	2559	2553	
	EPHA5-like	2602	2596	
	CLUH	2676	2670	
	SLC25A35	2786	2780	
	FAM222A	2843	2837	
	FAM222A	2176	2170	
GATA2			1822:1643	0.000004
	SLC17A6	1762	1757	
	KHDRBS2	1752	1747	
	KHDRBS2	1663	1658	
	PTGR2	1692	1687	
	SYNDIG1L	1745	1740	
	AREL1	1686	1681	
	YLPM1	1769	1764	
	YLPM1	1643	1638	
	PROX2	1822	1817	
	PROX2	1776	1771	
	RPS6KL1	1748	1743	
	MLH3	1784	1779	
	ACYP1	1694	1689	
	ZC2HC1C	1793	1788	
	ZC2HC1C	1662	1657	
	FLVCR2-like	1793	1788	
	IFT43	1712	1707	
	PDCL2	1811	1806	
	LOC102166866	1779	1774	
	EPHA5-like	1652	1647	
	CLUH	1814	1809	
	CLUH	1733	1728	
	RANGRF	1780	1775	
	RANGRF	1693	1688	
	ARHGEF15	1668	1663	
	ODF4	1791	1786	
	FAM222A	1778	1773	
ELK1			2130:1790	0.000005
	YLPM1	2070	2060	
	PROX2	2127	2117	
	RPS6KL1	2130	2120	
	RPS6KL1	2078	2068	
	RPS6KL1	2001	1991	
	RPS6KL1	1841	1831	
	PGF	1898	1888	
	PGF	1818	1808	
	MLH3	1922	1912	
	MLH3	1790	1780	
	ACYP1	1916	1906	
	JDP2	1887	1877	
	FLVCR2-like	1844	1834	
	IFT43	1864	1854	
	LOC102159670	1894	1884	

Table S6. Continue

	ZNF320-like	2052	2042	
	ZNF320-like	2040	2030	
	CLOCK	2085	2075	
	EPHA5-like	2086	2076	
	METTL16	1812	1802	
	LOC102165360	2067	2057	
	LOC102165360	2038	2028	
	PFAS	1951	1941	
	ODF4	1841	1831	
NKX3-1		371:49		0.000006
	SLC17A6	262	255	
	SLC17A6	149	142	
	SLC17A6	135	128	
	GCLC	245	238	
	KHDRBS2	49	42	
	SYNDIG1L	176	169	
	FCF1	56	49	
	PROX2	271	264	
	RPS6KL1	371	364	
	RPS6KL1	273	266	
	RPS6KL1	186	179	
	PGF	189	182	
	NEK9	304	297	
	IFT43	347	340	
	LOC102159670	350	343	
	TMEM165-like	206	199	
	PAFAH1B1	54	47	
	LOC102165360	337	330	
	LOC102165360	323	316	
	LOC102165360	256	249	
	LOC102165360	236	229	
	CCDC92-like	361	354	
	ODF4	235	228	
	FAM222A	117	110	
	FAM222A	96	89	
ARID3A		39:9		0.00001
	KHDRBS2	27	21	
	AREL1	39	33	
	PROX2	27	21	
	MLH3	15	9	
	ACYP1	9	3	
	LOC102166866	27	21	
	PAFAH1B1	32	26	
	PAFAH1B1	21	15	
	LOC102165360	14	8	
TFAP2A		3111:2988		0.00001
	GCLC	2988	2979	
	KHDRBS2	3069	3060	
	PTGR2	3110	3101	
	FCF1	3082	3073	

Table S6. Continue

	PROX2	3086	3077	
	PGF	3095	3086	
	ACYP1	3045	3036	
	ACYP1	2997	2988	
	ZC2HC1C	3111	3102	
	IFT43	3000	2991	
	LOC102165360	3093	3084	
	LOC102165360	3074	3065	
	CCDC92-like	3100	3091	
	CCDC92-like	2997	2988	
	SLC25A35	3046	3037	
	ARHGEF15	3104	3095	
	FAM222A	3063	3054	
	FAM222A	3003	2994	
SOX9		663:240		0.00002
	SLC17A6	525	516	
	PRIM2	659	650	
	PRIM2	611	602	
	VRTN	440	431	
	VRTN	289	280	
	SYNDIG1L	440	431	
	AREL1	541	532	
	AREL1	297	288	
	YLPM1	663	654	
	YLPM1	531	522	
	YLPM1	522	513	
	YLPM1	423	414	
	YLPM1	317	308	
	MLH3	378	369	
	ACYP1	631	622	
	NEK9	593	584	
	NEK9	317	308	
	TGFB3	272	263	
	IFT43	367	358	
	C8H4orf45	240	231	
	LOC102159670	535	526	
	ZNF320-like	589	580	
	ZNF320-like	256	247	
	NMU	438	429	
	CLOCK	644	635	
	CLOCK	497	488	
	LOC102165360	270	261	
	CLUH	515	506	
	CLUH	414	405	
	CCDC92-like	271	262	
	RANGRF	551	542	
	SLC25A35	574	565	
	ARHGEF15	373	364	
	FAM222A	340	331	

Table S6. Continue

FOXF2		211:75		0.00002
	KHDRBS2	133	119	
	PRIM2	141	127	
	AREL1	124	110	
	FCF1	149	135	
	FCF1	107	93	
	PROX2	174	160	
	RPS6KL1	148	134	
	RPS6KL1	92	78	
	PGF	75	61	
	MLH3	194	180	
	FLVCR2-like	178	164	
	TGFB3	185	171	
	C8H4orf45	143	129	
	C8H4orf45	125	111	
	LOC102166866	132	118	
	LOC102166866	79	65	
	TMEM165-like	184	170	
	EPHA5-like	87	73	
	METTL16	142	128	
	PAFAH1B1	211	197	
	LOC102165360	168	154	
	ARHGEF15	207	193	
FEV		1594:1514		0.00002
	KLHL31	1579	1571	
	AREL1	1572	1564	
	AREL1	1543	1535	
	FCF1	1514	1506	
	YLPM1	1527	1519	
	RPS6KL1	1594	1586	
	ACYP1	1588	1580	
	TGFB3	1581	1573	
	IFT43	1578	1570	
	IFT43	1545	1537	
	C8H4orf45	1587	1579	
	C8H4orf45	1520	1512	
	NMU	1557	1549	
	CLOCK	1552	1544	
	ARHGEF15	1559	1551	
GABPA		3035:2847		0.00003
	GCLC	3008	2997	
	GCLC	2997	2986	
	KHDRBS2	2993	2982	
	KHDRBS2	2896	2885	
	VRTN	2848	2837	
	RPS6KL1	2854	2843	
	PGF	2964	2953	
	ACYP1	2874	2863	
	NEK9	3023	3012	
	NEK9	2963	2952	

Table S6. Continue

	NEK9	2919	2908	
	FLVCR2-like	2856	2845	
	IFT43	2970	2959	
	C8H4orf45	2881	2870	
	LOC102159670	2917	2906	
	TMEM165-like	2847	2836	
	LOC102165360	2988	2977	
	LOC102165360	2923	2912	
	CCDC92-like	2895	2884	
	RANGRF	2926	2915	
	SLC25A35	3035	3024	
Myb		2163:1716		0.00003
	PTGR2	1752	1744	
	AREL1	1932	1924	
	PROX2	1761	1753	
	ACYP1	1927	1919	
	FLVCR2-like	2049	2041	
	TGFB3	1963	1955	
	TGFB3	1716	1708	
	RAPGEF2	2133	2125	
	ZNF320-like	2147	2139	
	CLOCK	2163	2155	
	CLOCK	1735	1727	
	SRD5A3	2090	2082	
	SRD5A3	1936	1928	
	EPHA5-like	2146	2138	
	METTL16	1982	1974	
	METTL16	1933	1925	
	PAFAH1B1	1787	1779	
	RANGRF	1883	1875	
	RANGRF	1841	1833	
	SLC25A35	1907	1899	
	ODF4	1960	1952	
	TRPV4	2006	1998	
	FAM222A	1903	1895	
PLAG1		3057:2948		0.00003
	KHDRBS2	3057	3043	
	PRIM2	2996	2982	
	VRTN	3028	3014	
	SYNDIG1L	2992	2978	
	RPS6KL1	3020	3006	
	ACYP1	3047	3033	
	ACYP1	2999	2985	
	ACYP1	2956	2942	
	NEK9	3052	3038	
	NEK9	2966	2952	
	FLVCR2-like	2967	2953	
	LOC102159670	3045	3031	
	ZNF320-like	3012	2998	
	ZNF320-like	2950	2936	

Table S6. Continue

	PAFAH1B1	2975	2961	
	LOC102165360	3039	3025	
	LOC102165360	2998	2984	
	LOC102165360	2954	2940	
	CCDC92-like	3057	3043	
	CCDC92-like	2990	2976	
	CCDC92-like	2948	2934	
	RANGRF	3017	3003	
	SLC25A35	2990	2976	
	SLC25A35	2975	2961	
	ARHGEF15	3020	3006	
	FAM222A	2982	2968	
	FAM222A	2949	2935	
Nkx2-5		175:20		0.00004
	KHDRBS2	45	38	
	KHDRBS2	37	30	
	KHDRBS2	29	22	
	SYNDIG1L	175	168	
	AREL1	73	66	
	AREL1	38	31	
	ACYP1	143	136	
	FLVCR2-like	119	112	
	TGFB3	47	40	
	PAFAH1B1	20	13	
	LOC102165360	147	140	
RORA_2		1125:1027		0.00005
	SYNDIG1L	1027	1013	
	YLPM1	1120	1106	
	RPS6KL1	1125	1111	
	PGF	1047	1033	
	MLH3	1122	1108	
	TGFB3	1112	1098	
	TGFB3	1051	1037	
	IFT43	1104	1090	
	NMU	1049	1035	
	PDCL2	1099	1085	
	PAFAH1B1	1112	1098	
	LOC102165360	1089	1075	
	CLUH	1054	1040	
	CCDC92-like	1033	1019	
	PFAS	1047	1033	
	ODF4	1045	1031	
Mycn		3123:2508		0.00005
	GCLC	3113	3103	
	GCLC	2652	2642	
	KHDRBS2	2542	2532	
	PRIM2	2592	2582	
	PRIM2	2570	2560	
	SYNDIG1L	2728	2718	
	SYNDIG1L	2508	2498	

Table S6. Continue

	FCF1	3090	3080	
	FCF1	2982	2972	
	YLPM1	2955	2945	
	PROX2	2930	2920	
	RPS6KL1	3121	3111	
	MLH3	3123	3113	
	ACYP1	2972	2962	
	ACYP1	2930	2920	
	ZC2HC1C	2915	2905	
	NEK9	2574	2564	
	RAPGEF2	2514	2504	
	EPHA5-like	2758	2748	
	EPHA5-like	2604	2594	
	LOC102165360	3094	3084	
	CLUH	2678	2668	
	CCDC92-like	2961	2951	
	CCDC92-like	2899	2889	
	SLC25A35	3122	3112	
	FAM222A	2845	2835	
	FAM222A	2767	2757	
	FAM222A	2687	2677	
	FAM222A	2666	2656	
SP1		3108:3070		0.00006
	KHDRBS2	3103	3093	
	SYNDIG1L	3093	3083	
	SYNDIG1L	3075	3065	
	AREL1	3077	3067	
	PROX2	3091	3081	
	RPS6KL1	3091	3081	
	PGF	3108	3098	
	JDP2	3083	3073	
	JDP2	3071	3061	
	FLVCR2-like	3087	3077	
	FLVCR2-like	3076	3066	
	IFT43	3108	3098	
	IFT43	3097	3087	
	IFT43	3081	3071	
	ZNF320-like	3108	3098	
	ZNF320-like	3070	3060	
	NMU	3072	3062	
	SLC25A35	3101	3091	
	TRPV4	3090	3080	
ELF5		1599:1546		0.00006
	KLHL31	1599	1590	
	KLHL31	1580	1571	
	PTGR2	1553	1544	
	VRTN	1574	1565	
	AREL1	1591	1582	
	JDP2	1570	1561	
	TGFB3	1581	1572	

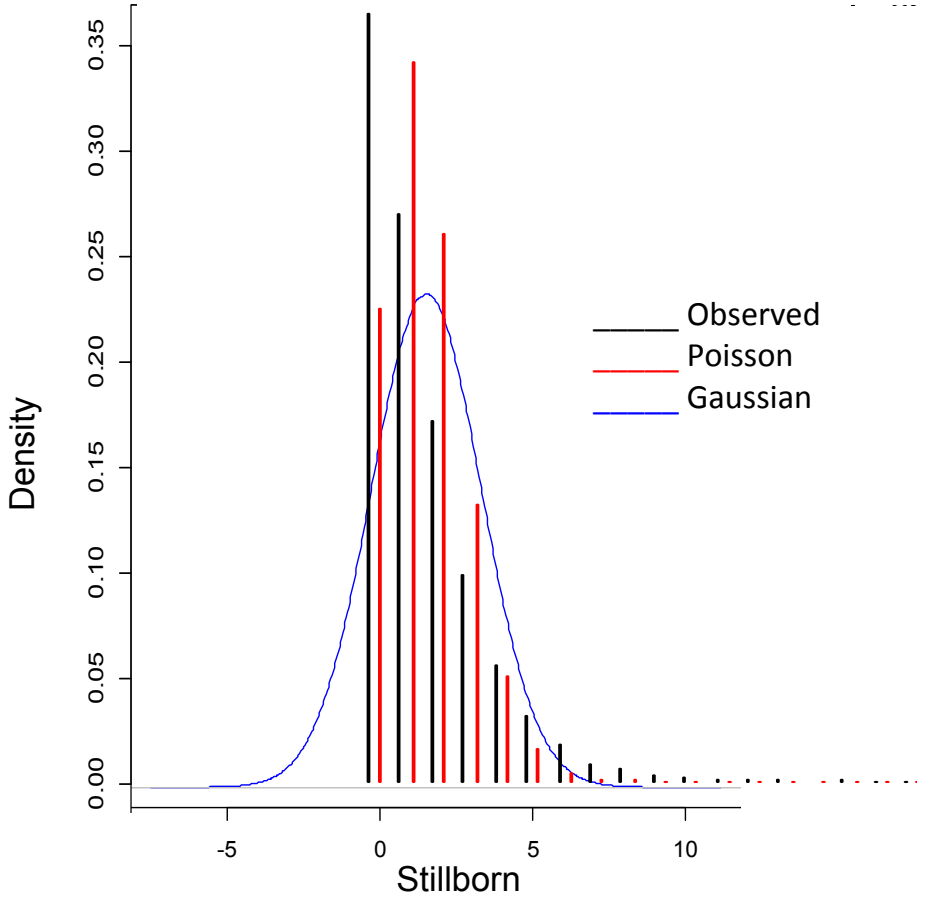
Table S6. Continue

	IFT43	1578	1569	
	IFT43	1546	1537	
	C8H4orf45	1583	1574	
	NMU	1558	1549	
	CLUH	1559	1550	
	ARHGEF15	1560	1551	
Pax2		1036:696		0.00007
	KLHL31	1000	992	
	PRIM2	1008	1000	
	PRIM2	990	982	
	PTGR2	774	766	
	AREL1	975	967	
	AREL1	876	868	
	AREL1	755	747	
	YLPM1	717	709	
	PROX2	1036	1028	
	PROX2	723	715	
	RPS6KL1	755	747	
	PGF	925	917	
	ZC2HC1C	888	880	
	JDP2	959	951	
	JDP2	766	758	
	ZNF320-like	867	859	
	LOC102166866	885	877	
	TMEM165-like	702	694	
	LOC102165360	1025	1017	
	CLUH	696	688	
	CCDC92-like	722	714	
	PFAS	837	829	
	SLC25A35	770	762	
	ARHGEF15	940	932	
RXRA::VDR		146:111		0.00007
	KHDRBS2	127	112	
	PRIM2	111	96	
	PGF	114	99	
	MLH3	146	131	
	TGFB3	146	131	
	PDCL2	127	112	
	TMEM165-like	124	109	
	PAFAH1B1	111	96	
ELF5		1179:1097		0.00007
	GCLC	1170	1161	
	KHDRBS2	1145	1136	
	SYNDIG1L	1130	1121	
	AREL1	1116	1107	
	YLPM1	1155	1146	
	RPS6KL1	1179	1170	
	RPS6KL1	1133	1124	
	PGF	1128	1119	
	ACYP1	1148	1139	

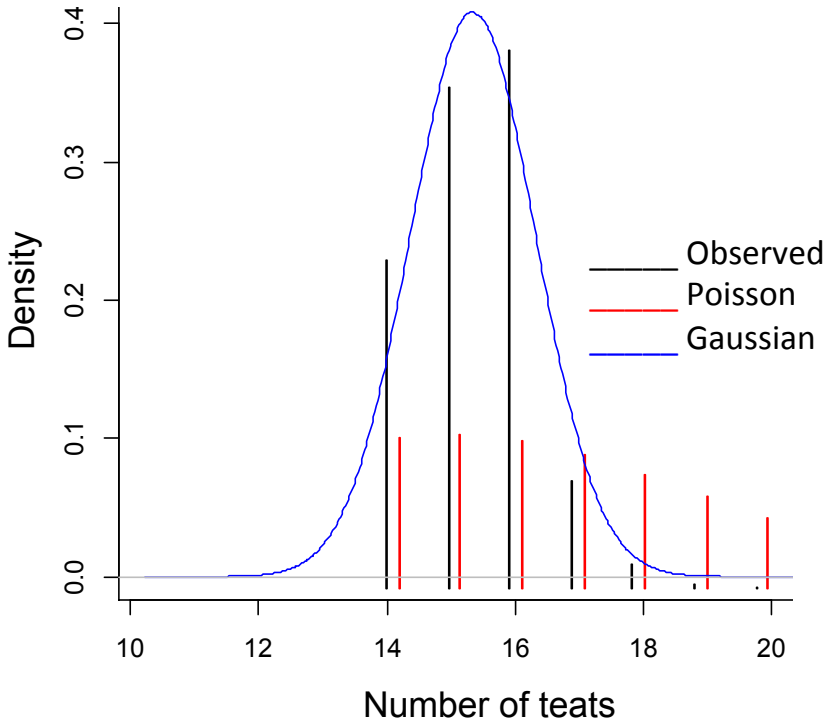
Table S6. Continue

NEK9	1116	1107
FLVCR2-like	1102	1093
C8H4orf45	1107	1098
NMU	1139	1130
CLOCK	1176	1167
TMEM165-like	1176	1167
EPHA5-like	1104	1095
CCDC92-like	1167	1158
TRPV4	1097	1088

*Genes are repeated according to the number transcription factor binding sites



Additional file 9: Figure S3 Density of stillborn (SB) data considering the observed, Poisson and Gaussian curves.



Additional file 10: Figure S4 Density of number of teats (NT) data considering the observed, Poisson and Gaussian curves.

Chapter 4

Post-GWAS: gene networks elucidating candidate genes divergences for number of teats across two pig populations

Lucas L Verardo[1, 2*], Marcos S Lopes[2, 3], Susan Wijga[3], Ole Madsen[2], Fabyano F Silva[1], Martien AM Groenen[2], Egbert F Knol[3], Paulo S Lopes[1], Simone EF Guimarães[1]

[1]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, 36570000, Brazil

[2] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6700 AH, the Netherlands

[3] Topigs Norsvin Research Center, Beuningen, 6641 SZ, the Netherlands

BMC Genomics (Submitted)

Abstract

Background

Number of teats (NT) is an important trait affecting both, piglet's welfare and production level of pig farms. Biologically, embryonic mammary gland development requires the coordination of large number of signalling pathways necessary for a proper teats morphogenesis. Several QTLs for NT have been identified, however, further analysis is still lacking. Gene networks derived from significant SNPs of a GWAS can be used to examine the process of shared pathways and functions of genes allowing the identification of the most likely candidates for the trait based on their biological connection in the gene network. Besides, such analyses may also be helpful for understanding the genetic diversity between populations for the same trait(s). Thus, significant SNPs were identified separately for a Landrace-based (C) and Large White-based (D) dam lines. Gene-transcription factors networks was construct aiming to obtain the most likely candidate genes for NT in each line, followed by a comparative analysis between both lines to access (dis)similarities at marker and gene level.

Results

From separate GWAS analyses on the two lines, 24 and 19 significant SNPs for NT were identified for line C and D, respectively. Only one significant SNP was common to both lines. Each significant SNP set was used to identify underlying genes resulting in 31 and 23 genes for line C and D, respectively. Only three genes were common to both lines. However, comparative analysis on gene ontology, known transcription factors (TF) binding sites, and literature review revealed relevant common biological processes for NT, including mammary gland and spinal cord development.

Conclusions

Our network analysis illustrated genes interactions consistent with known mammal's breast biology and captured known TF. We observed different sets of candidate genes for NT in each of the lines evaluated that may have an effect on the phenotype. This

genomic diversity across lines should be taken into account in breeding programs in such a way that breed specific reference populations should be considered for genomic selection. Moreover, complex traits, such as NT, are subject to the interaction of large number of genes regulated by a TFs variety; many of them still to be identified.

Keywords

Bayesian GWAS, biological process, complex traits

Background

In the last decades, breeding programs have achieved significant improvement in litter size. As a consequence, the number of piglets in a litter is often larger than the number of teats (NT) of the sow [1]. Lower NT increases suckling competition, which can result in higher mortality before weaning. Therefore, NT is directly related to the mothering ability of sows [2] and it may affect both welfare of piglets and production level of pig farms. Thus, a better understanding of the genetic background of NT would provide the opportunity of performing more efficient selection.

Biologically, embryonic mammary gland development requires the coordination of a large number of signaling pathways to direct the cell shape changes, cell movements, and cell–cell interactions necessary for proper morphogenesis [3]. During prenatal life, at histological level, distinct bands in the surface ectoderm, so called mammary or milk lines, connects all positions where mammary glands may form on either side of the body common in any given mammalian species [4], [5]. In addition, units of paraxial mesoderm, the somites, are among others related to the formation of vertebrae and ribs [6] that could also have a direct relation with the final NT observed in pigs [7].

Studies for NT have been performed on different populations using microsatellites markers [2], [8], [9]. In general, these studies identified quantitative trait locus (QTL) regions with a large confidence interval due to the low number of markers. With the development of dense maps of Single Nucleotide Polymorphism markers (SNPs),

several genome-wide association studies (GWAS) have been performed for NT [10], [11], which allowed narrowing down the confidence interval of previously identified QTLs and also the identification of novel QTLs. However, a further analysis of these QTLs with post-GWAS analysis is still lacking. The further genetic dissection of QTLs for complex phenotypes can be accomplished through the analysis of candidate gene networks.

Gene networks derived from significant SNPs of a GWAS can be used to examine the process of shared pathways and functions across these genes hereby allowing the identification of the most likely candidate genes based on their biological connection in the gene network. In cattle, gene networks for puberty-related traits have been performed [12], [13] and in pigs this approach has also begun to be exploited [14]. In addition to gene networks, transcription factors (TF) analyses can be performed. It has been shown that TF play a role in important traits in pigs such as e.g. *PIT1* for carcass traits and *SREBF1* for muscle fat deposition [15], [16]. Evidence for the interaction between a TF (playing a role in a trait) and its predicted targets can help in pointing out the most probable group of candidate genes for the studied trait through a gene-TF network.

Besides the biological information provided by gene-TF networks, such analyses may also be helpful for understanding genetic diversity between populations for the same trait(s). As distinct populations tend to have specific QTL for the same trait [17], information from significant regions in each population may illustrate the similarities and connections of distinct genes through TFs. These analyses have the potential to clarify the finding of different sets of genes for the same trait not only across populations, but also across time (e.g. tick resistance in cattle across dry and rainy season) [18] and environment (e.g. sire evaluations across environments in pigs) [19].

Toward this aim, we present a gene network analysis across two dam lines based on a Bayesian GWAS for NT. For each line separately, genes linked to significant SNPs were identified. These genes were used to construct a combined network and also to detect related TF in each line, aiming to obtain the most likely candidate genes for NT in each line, followed by a comparative analysis to assess (dis)similarities at marker and gene level across the lines.

Methods

Ethics Statement

The data used for this study was obtained as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purposes of the breeding program. Data recording and sample collection were conducted strictly in accordance with the Dutch law on the protection of animals (Gezondheids- en Welzijnswet voor Dieren).

Data

A total of 3,983 animals were genotyped using the Illumina PorcineSNP60 BeadChip [20]. These animals were originated from two dam lines: line C (Landrace-based, n=1,913) and line D (Large White-based, n=2,070). Positions of the SNPs were based on pig genome build10.2 [21]. Cleaning of the SNP data consisted of exclusion of SNPs with: 1) a GenCall score <0.15 and located on sex chromosomes, 2) a call rate <95%, 3) a minor allele frequency (MAF) <0.01 and 4) no physical position on pig genome build 10.2. After these quality control measures, out of 64,232 SNPs, 37,412 and 37,782 SNPs respectively, remained for lines C and D for the GWAS. All animals had frequencies of missing genotypes <0.05.

Genome-wide association analysis

Estimated breeding values (EBVs) for NT were obtained from the genetic evaluation routine used by the breeding company Topigs Norsvin using MiXBLUP [22]. The

model for obtaining the EBV for NT included the fixed effects of herd-year-season, sex and line, and the additive genetic effect (animal) as a random effect. Reliabilities per animal were extracted from the genetic evaluation and were based on the methodology of Tier and Meyer [23]. The EBVs were deregressed using the methodology proposed by Garrick et al. [24].

A Bayesian Variable Selection model [25] was fitted for NT by estimating the marker effects with all SNPs simultaneously by using the following model:

$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta}$,

where $\mathbf{y}$ is the vector of deregressed EBVs (on n animals), $\boldsymbol{\mu}$ is a vector equal to the mean, $\mathbf{X}$ of dimension n by p matrix, p being SNP genotypes coded as 0, 1, or 2 copies of a particular allele and $\boldsymbol{\beta}$ is a vector (of length p SNP) with the marker effects. The assumed distribution of this vector was: $\boldsymbol{\beta} \sim \begin{cases} \mathrm{N}(0, \sigma^2_{g_0}), \text{ with probability } \pi_0, \\ \mathrm{N}(0, \sigma^2_{g_1}), \text{ with probability } \pi_1 = 1 - \pi_0, \end{cases}$

where the first distribution refers to the SNPs which are assumed to have a small effect and the second distribution contains SNPs assumed to have a large effect. The probability to be in the first distribution (π_0) was set to 0.999, meaning only 1 in every 1000 SNPs are expected to have a large effect, which is on average 42 SNP per cycle. The term *e* is a vector of length n with random residual effects assumed to be normally distributed but weighted, $\mathrm{N}(0, \sigma^2_e\mathbf{W})$, where $\mathbf{W}$ is a diagonal matrix with elements $w_1, \ldots, w_n$. The model was implemented in the program Bayz (https://www.bayz.biz).

A total of 500,000 Markov chain Monte Carlo (MCMC) samples chains with a burn-in of 5,000 cycles were run and a Metropolis-Hastings sampler was applied to get good convergence which was assessed by visual inspection of the trace and using Gelman and Rubin's convergence diagnostic based on deviance [26] using the R package CODA [27].

To determine which SNPs are significantly associated, a Bayes Factor (BF) was calculated for every SNP using the prior probability (π_0 and π_1) and the posterior probability ($\hat{p}_i$) by calculating an odds ratio as:

$$BF = \frac{\hat{p}_i/(1-\hat{p}_i)}{\pi_1/\pi_0}$$

SNPs with a BF ≥ 100 were considered to have a significant association with NT, cited as 'decisive' according to Kass and Raftery [28]. Linkage disequilibrium (LD) relations among the significant SNPs were evaluated through Haploview [29] to identify QTL regions (blocks) on chromosomes using the default parameters based on Gabriel et al. [30]. A 95% confidence boundary on D prime [31] was used to determine if a pair of SNPs was in "strong LD". Markers with MAF < 0.01 were ignored.

Network analyses

Genes overlapping single markers plus 32 Kbp 5' and 3' flanking sequence, (e.i. the half of the average distance between two SNPs on the Chip) were identified at the reference genome (build10.2). By including the 32 Kbp intervals, we verified the presence of any gene that could be linked with a QTL or a significant SNP for NT. The GeneCards web site (http://www.genecards.org/) and the program TOPPCLUSTER (http://toppcluster.cchmc.org/) were used to obtain the functional Gene Ontology (GO) terms and pathways corresponding to the identified genes through the human orthologs. The ClueGO Cytoscape plugging [32] was used to construct a gene network highlighting the biological roles and relations across both set of genes from dam lines C and D at the same time.

Aiming to identify the TF related to the identified genes for each line, the TFM-Explorer program (http://bioinfo.lifl.fr/TFM/TFME/) was used. This program takes a set of gene sequences, and searches for locally overrepresented transcription factor binding sites (TFBS) using weight matrices from the JASPAR vertebrate database [33] to detect all potential TFBS, and extracts significant clusters (region of the input sequences

associated with a factor) by calculating a score function. This score threshold is chosen to give a P-value equal or better to 10^{-2} for each position for each sequence such as described in Touzet and Varré [34]. From each set of genes, excluding ncRNA genes, we collected the 3000 bp upstream and 300 bp downstream flanking sequence from the gene start site based on Sscrofa10.2 assembly at NCBI web site. Data from each line was separately used as input at TFM-explorer and the given list of TF was submitted to Cytoscape [35] using a Biological Networks Gene Ontology tool (BiNGO) plugin [36], to determine which gene ontology (GO) terms are significantly overrepresented. Based on biological processes (e.g. mammary gland development) and a literature review, we were able to select the main TFs related to NT (key TF) in each population, from which we constructed a gene-TF network.

Aiming to identify the most likely candidate genes, we used the NetworkAnalyzer plugin within Cytoscape. Based on the number of TFBS and consequently the number of connections in each node, the most connected genes within the NT gene-TF network were determined. Finally, the gene-TF networks derived for both lines were merged, highlighting the genes and TF in common providing an overview across the two dam lines.

Results

Genome-wide association analysis

A total of 24 and 19 significant SNPs for NT were identified for lines C and D, respectively. These significant SNPs were spread across 12 and 8 chromosomes in the lines C and D, respectively (Table 1). Only one SNP in common between the two lines (rs81396056, at chromosome 7) was identified. Using the significant SNPs we found 31 and 23 genes for lines C and D, respectively (Table 1). Three genes (*NPC2*, *LTBP2* and *ISCA2)* were shared by the two lines due to marker rs81396056.

Table 1 - Significant SNPs and linked genes. Significant SNPs, their positions in base pairs (bp) at swine chromosome (Chr), their belonged dam Line, and the associated gene (located in an interval of 32 Kbp around each SNP) followed by their distance in bp from the marker.

SNP	Chr	Position (bp)	Line	Gene	Distance (bp)
ASGA0096071	2	61257750	C	LOC100525329/OR10H3-like	18205
				LOC100525507/OR10H3-like	1517
				LOC100525684/OR10H4-like	20966
				LOC100525859/OR10H4-like	27309
ALGA0013914	2	62015620	C	LOC100628051/OR10T2-like	24458
				ILVBL	7571
				SYDE1	726
				LOC100523537/OR1I1-like	15768
ASGA0094490	3	125568713	C	GDF7	15598
				LOC102159651/WBP11-like	12824
				HS1BP3	895
ASGA0019959	4	71146044	C	LOC100738208/SLCO5A1-like	18364
				LOC100738252	in
ALGA0025515	4	71218178	C	LOC100738252	in
				SLCO5A1	5929
MARC0020649	6	5346336	C	CDH13	in
ASGA0029386	6	120405392	C	-	-
MARC0035827	6	122355118	C	-	-
DIAS0000810	7	26954317	C	DHX16	24300
				PPP1R18	9953
				NRM	6869
				MDC1	in
				TUBB2A	15390
				FLOT1	20807
H3GA0021033	7	40029052	D	KIF6	in
ALGA0040656	7	40465318	D	MOCS1	14088
H3GA0022644	7	102901720	D	PTGR2	27013
MARC0038565	7	103495170	D	VRTN	28094
				SYNDIG1L	5675
ASGA0035500	7	103574383	C and D	NPC2	in
				ISCA2	8350
				LTBP2	14820
DIAS0000795	7	103594753	C	NPC2	12395
				ISCA2	5566
				LTBP2	in
M1GA0010658	7	103999954	D	RPS6KL1	24266
				PGF	in
ALGA0043983	7	104352654	D	-	-
ALGA0122954	7	104598913	D	JDP2	in
				LOC100624918/FLVCR2-like	8951

Table 1 (continue)

ALGA0109821	8	41799445	D	RASL11B	30214
MARC0097287	8	80774753	D	ARFIP1	in
ASGA0039113	8	84325429	C	LOC102165402	in
ASGA0039115	8	84361242	C	LOC102165402	in
ALGA0052334	9	36148167	D	PGR LOC102165942	9627 in
MARC0087389	9	123847767	C	-	-
ALGA0103761	10	52584711	D	FRMD4A	in
MARC0063711	10	52587829	D	FRMD4A	in
MARC0018399	10	52679135	D	FRMD4A	in
ASGA0048983	10	72513357	C	-	-
H3GA0055392	10	72941077	C	LOC100624137	285
ASGA0052010	11	81585578	D	EFNB2	in
ASGA0054746	12	44762306	D	CRLF3	in
ALGA0118253	12	53322741	C	-	-
ASGA0071205	15	143040920	C	CCL20	9408
ASGA0071569	15	148722557	C	TRPM8 SPP2	19985 4738
ALGA0106798	15	149411138	D	-	-
DRGA0016094	16	35884503	C	SNX18	25517
ASGA0078927	18	13918533	D	-	-
ALGA0096935	18	8952443	C	-	-
ASGA0098911	18	49576732	C	-	-
DRGA0017043	18	49945197	C	EVX1	15640
ASGA0092614	18	51077232	C	-	-

Network analyses

Information about biological processes, cellular components and molecular functions of the identified genes, based on human orthologs, was collected (Additional file 1: Table S1 and Additional file 2: Table S2). Further, the gene network highlighted their related biological processes such as mammary gland duct morphogenesis and cell differentiation in spinal cord (Figure 1 and Additional file 3: Figure S1). In addition, regulatory sequence analyses were performed for all detected genes identifying transcription factors (TF) significantly enriched (P-value<0.001) (Additional file 4: Table S3 and Additional file 5: Table S4) resulting in 22 and 21 TF for lines C and D set of genes respectively. The main TFs associated with NT based on biological

processes and a literature review (Table 2) were chosen to generate a gene-TF network for each line. Based on the separate analyses, a merged network was constructed (Figure 2) which enabled us to identify new candidate genes for NT in each line (e.g., *CDH13, EVX1* and *GDF7* on line C and *EFNB2, KIF*6 and *SUZ12* on line D).

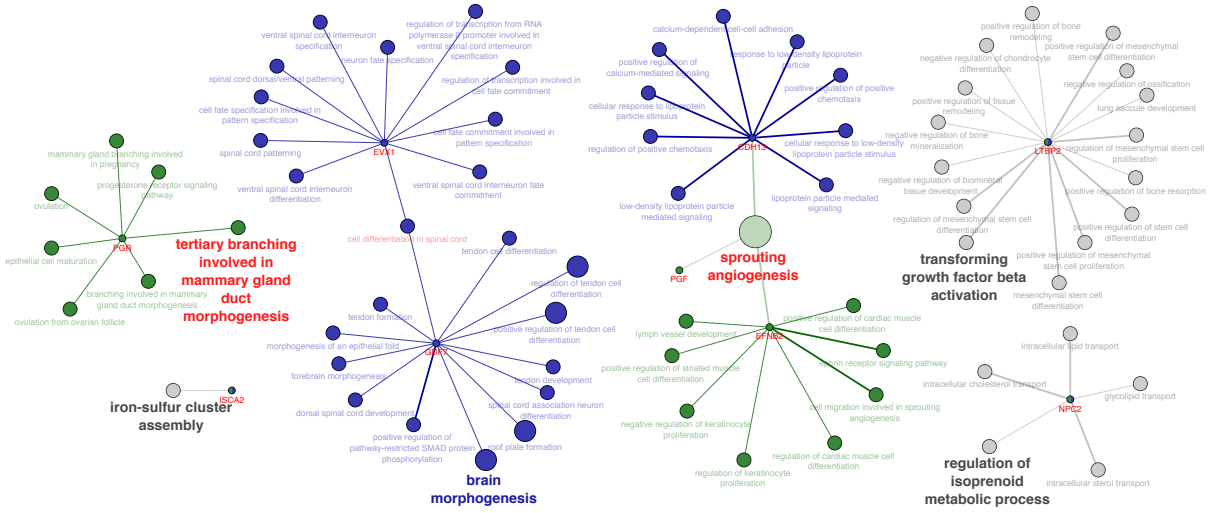


Figure 1 – Main biological process networks. Functionally group networks with biological process terms and genes (labeled in red) as nodes. Blue nodes represent line C while green nodes line D. The gray nodes are related with genes shared between both lines. The node size represents the term enrichment significance from ClueGO. The most significant term per group is shown in bold and the most related with the trait are labeled in red.

Table 2 – Main Transcription Factors. Most representative transcription factors (TF), associated with genes identified in lines C and D, based on their biological process and literature review

TF	Line	Biological Process	Literature review
E2F1	C	Central nervous system development	Endochondral Bone Formation (Scheijen et al., 2003)
FOXA2	C	Central nervous system development	Formation of the intervertebral Discs (Maier et al., 2013)
RUNX1	C and D	Regulation of granulocyte differentiation	Mammary ductal growth (Wall et al., 2014) and Skeletal development (Smith et al., 2005)
KLF4	C and D	Cell fate commitment involved in the formation of primary germ layers	Mammary epithelial cells (Miller et al., 2001)
TFAP2A	C and D	Tissue development	Developing fetal human breast (Friedrichs et al., 2007)
EBF1	D	Developmental process	Somite development (El-Magd et al., 2014)
ELF5	D	Mammary gland development	Mammary gland tissue (Choi et al., 2009)

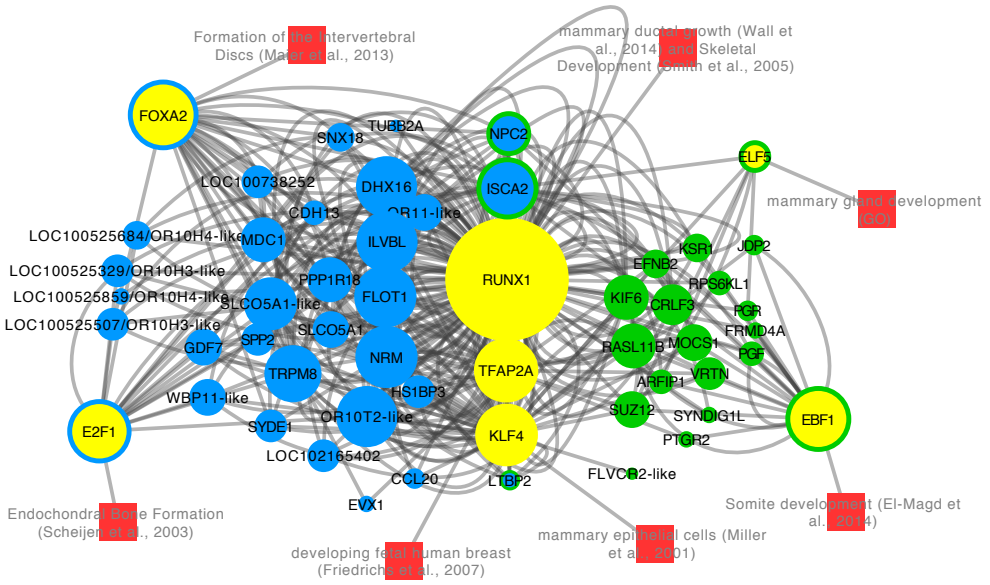


Figure 2 - Gene-Transcription Factor (TF) network. Genes related with significant SNPs for number of teats in each line (blue nodes, line C; green nodes, line D) and in common to both lines (blue nodes with green border). Associated with these genes, we have TFs common for both lines (yellow nodes), and specifics for each line (yellow node with blue border, line C and yellow nodes with green border, line D). The node size corresponds to the network analyses (Cytoscape) score where bigger nodes represent higher edges density associated with the number of TF binding sites. Red square nodes are the TF related biological process (GO term) and role based on literature review

As we observed three genes in common across the two lines (*NPC2*, *LTBP2* and *ISCA2)*, we performed the same analyses excluding them to assess their role in the final network connections. One of the three commons TFs (*RUNX1*) was still common between the two dam lines whereas the two other common TFs (*TFPA2A* and *KLF4*) were only identified in line C at this time (Figure 3). In addition new TFs, related with mammary gland and number of vertebrae in both lines (*PAX5* and *SOX2* for line C and *FOXC1* for line D) were now identified. The final set of genes present in the gene-TF network with and without overlapping genes was conserved.

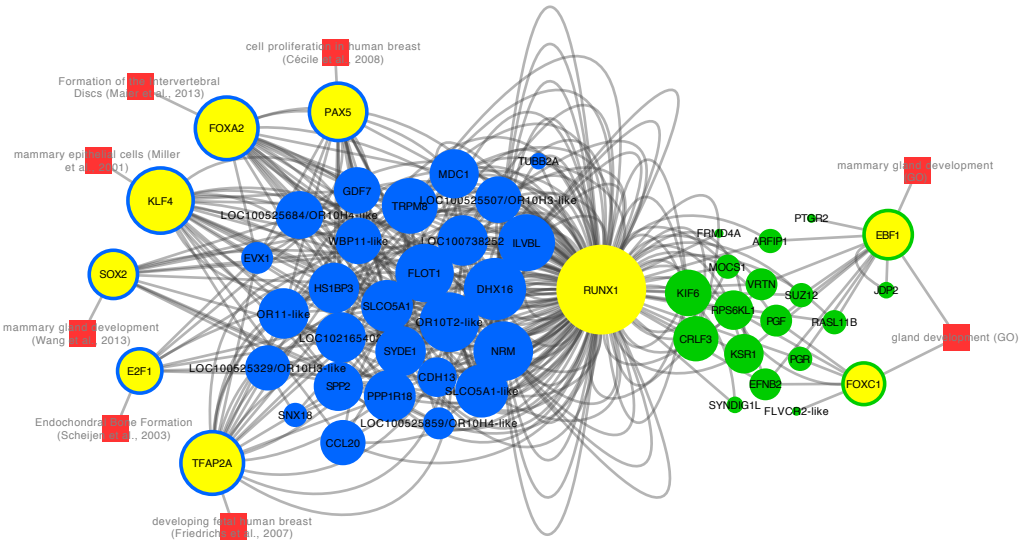


Figure 3 - Gene-Transcription Factor (TF) network excluding the three common genes. Genes related with significant SNPs for number of teats in each line (blue nodes, line C; green nodes, line D). Associated with these genes, we have TFs common for both lines (yellow nodes), and specifics for each line (yellow node with blue border, line C and yellow nodes with green border, line D). The node size corresponds to the network analyses (Cytoscape) score where bigger nodes represent higher edges density associated with the number of TF binding sites. Red square nodes are the TF related biological process (GO term) and roles based on literature review.

Discussion

Genome-wide association analysis

We used a Bayesian GWAS to identify significant SNPs for NT in two dam lines. Dam line C had 24 significant SNPs distributed across 12 different chromosomes. In other association studies using commercial pig lines [8]-[10], [37]-[40], several QTL for NT were reported at the same chromosome regions. For dam line D the 19 significant SNPs were distributed on 8 different chromosomes. Previous studies [2], [9], [10], [39], [40], have reported QTL for NT in commercial lines at the same chromosome regions identified in this study except for SSC 11 and 15.

From all significant SNPs identified in both lines, only one marker (rs81396056) was in common. In general, such lack of overlap across different populations has also been observed in other studies on other species and traits [18], [19]. More specific to animal breeding, this finding demonstrates the need for a biological understanding of complex traits for animal selection across populations and association studies. Thus, these findings highlight the importance of a post GWAS analyses such as gene-TF networks to elucidate genetic divergences.

Network analyses

From the significant SNPs for each line, 31 and 23 genes were identified for dam lines C and D, respectively. Three genes (*NPC2*, *LTBP2* and *ISCA2*) are in common between the two lines due to the shared marker (rs81396056*)* on chromosome 7. The *NPC2* gene encodes for a protein also called HE1, which it is a small soluble glyco-protein of the late endosomes [41], [42]. *LTBP2* encodes for a latent transforming growth factor beta binding protein 2, that has been associated with elastic fibers in developing elastic tissues [43] and *ISCA2* codes for an Iron-Sulfur Cluster Assembly 2 protein, acting in the mitochondria in assembling Iron and Sulfur [44]. Of these

three, the *LTBP2* gene is the most likely to affect the number of teats. This gene is associated with development of elastic tissues biological process, being important for mammary gland development.

To identify genes with a biological role in NT, their GO terms and pathways were collected and a network based on their biological process constructed. From this network it is possible to identify genes involved in the same biological process, but significantly associated to different significant markers and genomic regions from different populations, like *CDH13* in line C and *PGF* and *EFNB2* in line D both involved in sprouting angiogenesis. *CDH13*, also known as T-cadherin, together with adiponectin, has been found related to angiogenesis in mouse mammary gland [45]. *PGF* and *EFNB2* had been linked to hematopoiesis and angiogenesis [46], [47] suggesting that these genes could play important roles in mammary gland development and be strong candidate genes for NT.

Through the biological process network we could also identify genes involved with cell differentiation of spinal cord (*GDF7* and *EVX1;* line C) and a gene involved with mammary gland duct morphogenesis (*PGR*; line D). *GDF7* is a growth differentiation factor that has been found related with bone structural integrity [48] and *EVX1* to be expressed in the developing spinal cord [49]. At the same time, *PGR* gene is the progesterone receptor coding protein cited by Tanos et al. [50] to be involved in mammary gland development.

The set of genes from each line was used to explore promoter regions for enriched TFs from which the most representative (in terms of relevance for NT) from each line were selected to generate a gene-TF network. These networks were merged enabling visualization of the genes and TF in common between lines C and D. Three TF were in common across the two lines (*RUNX1, TFAP2A,* and *KLF4*). Of these three TF,

RUNX1 is the most connected (linked with a high number of genes) in the network. *RUNX1* is a runt-related transcription factor associated with skeletal development [51] and in mammary ductal growth [52]. Of the two other TFs, *KLF4* [53] and *TFAP2A* [54] have been associated to mammary epithelial cells and developing fetal human breast respectively.

There are four other well-connected TFs in the gene-TF network (*FOXA2* and *E2F1* in line C; and *EBF1* and *ELF5* in line D). *FOXA2* has been found required for formation of the intervertebral discs [55], an important process for the final body length. *E2F1* has been found to be involved with endochondral bone formation [56]. Of the two specific TFs for line D, *EBF1* is associated to somite development [57]. Somites are units of paraxial mesoderm related to, among others, the formation of vertebrae and ribs [6]. The last TF, *ELF5* is involved with the mammary gland tissue [58] and was also identified in a previous association study in line D using gene-TF network for number of teats (unpublished observations).

Taken together, from the gene-TF network we found candidate genes from each line for NT based on their TF binding sites network. For line C, besides *EVX1* and *GDF7* involved with spinal cord developing and bone structural integrity respectively, we also observed a number of genes well connected with the common set of TF (e.g. *WBP11, MDC1, OR10T2, ILVBL, TRPM8, DHX16, SLCO5A1, NRM, LOC100738252, LOC102165402, CCL20, OR1I1, FLOT1, SYDE1, PPP1R18, HS1BP3, TUBB2A* and *LOC100525859*) involved in mammary gland and skeletal development. From this group of genes, the *WBP11* and *MDC1* are target genes of tumor protein p63 [59], [60] which is involved in the adhesion program of mammary epithelial cells [61]. The *TRPM8* gene has been found expressed in the spinal cord of mice [62], and *DHX16* is involved in embryonic development [63].

As for line C, a number of genes in line D were well connected to the same set of TF. Being well connected with TFs involved with mammary gland development and paraxial mesoderm morphogenesis, the following genes were highlighted: *MOCS1, KIF6, PS6KL1, RASL11B, VRTN, CRLF3, EFNB2, SUZ12, ARFIP1, SYNDIG1L, PGR, FRMD4A, FLVCR2, PGF, PTGR2, KSR1* and *JDP2*. Besides the PGR gene that is related with mammary gland (observed through the biological process analyses), *VRTN* and *SYNDIG1L* has been suggested as candidate genes for the number of vertebrae and consequently related with number of teats [10], [11] while *KIF6* and *SUZ12* have been found to be involved with spine development and neural tube formation respectively [64], [65].

To test that the common TFs across both lines were not due to the three common genes (*NPC2*, *LTBP2* and *ISCA2*), we performed the same analyses excluding them from the analyses. Of the three common TFs, *RUNX1* remained when excluding the overlapping genes. Two TF that were previously in common between the two set of genes (*TFPA2A* and *KLF4*) were now only identified in line C together with some new enriched TFs: *PAX5* and *SOX2*. *PAX5* and *SOX2* have been found to be involved with mammary gland epithelial cells and development [66], [67]. For line D TF set, *FOXC1* was new in these analyses, being involved somitogenesis in zebrafish [68]. The final set of genes present at the gene-TF network was conserved between the two analyses. These analyses enabled to confirm that even without genes in common for the same trait across both lines; each set of genes are still linked through common TF and biological roles in mammary gland and vertebrate development.

Interestingly, as noted for each line, in general a diverse set of genes was identified but overlapping biological roles regarding NT (e.g. mammary gland development and spinal cord/vertebrae development). Thus, for the two populations/lines selection for

the same trait has led to selection of different genes involved in the same biological processes. The explanation for selection on different set of genes is still not fully understood but recent publications have found the same observation. Machado et al. [18] in a QTL study for tick resistance/susceptibility in a bovine F2 population found that depending on the tick evaluation season (dry and rainy), different sets of genes could be involved in the tick resistance mechanism. Veroneze et al. [17], observed differences in linkage disequilibrium patterns and persistence of phase across pig lines suggesting different QTL and distinct set of genes. Furthermore, in a genotype and environment interaction study, Silva et al. [19] observed variation in the significant SNP set for total number of piglets born in a commercial pig line according to environment variations resulting in different QTLs.

The gene-TF networks analysis in this study, further illustrate the phenomenon observed in previous studies, showing that the same trait can be under selection without selecting for the same set of genes. Selection can act on many genes and depending on the population (variants present in that) initial selection can be on different genes and this may directly have an influence on what genes are under selection at later stages. Instead, different genes involved in the same biological pathways may result in variation of a given trait and can be incidentally selected with the same phenotypic outcome. In a broad point of view, studies also have shown that many different genotypes can bring forth the same phenotypes [69], [70] as well as the gene networks showed for NT trait in the present study.

Different components of pig physiology may be important for determining the final number of teats [3]-[7], and different genes are therefore important and contribute to the observed genetic differences. Therefore, analyses of data from different lines or populations are very valuable for animal breeding as it provides additional insights

into the genes controlling the trait of interest under distinct circumstances. From our study, we conclude that within the breeding programs of these two different commercial lines, specific sets of genes have been selected for in each line explaining the observed different set of genes involved in the same trait phenotype. Therefore, more studies are needed to verify the role of these genes and the gene-TF network interactions in each population.

Conclusion

The present study provided a rich information resource about genes identified in a Bayesian GWAS for NT in two pig lines. The network analysis within and across the two dam lines illustrated genes interactions that were consistent with the known mammal's breast biology and captured known TF, pointing out different sets of candidate genes for NT for each of the lines evaluated (*WBP11, MDC1, TRPM8, DHX16, CDH13, EVX1* and *GDF7* for line C and *EFNB2, KIF6, SUZ12, VRTN* and *SYNDIG1L* for line D). Based on these results, the genomic diversity across lines should be taken into account in breeding programs in such a way that breed specific reference populations should be considered for genomic selection. Moreover, complex traits, such as number of teats, are subject to the interaction of a large number of genes regulated by a variety of transcription factors; many of them still to be identified.

Abbreviations

NT: number of teats
QTL: Quantitative Trait Locus
SNP: Single Nucleotide Polymorphisms
GWAS: Genome-wide association studies
TF: transcription factors
MAF: minor allele frequency
EBV: estimated breeding values
MCMC: Markov chain Monte Carlo
BF: Bayes Factor
LD: Linkage disequilibrium

TFBS: transcription factor binding sites
GO: gene ontology

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LLV, FFS, EFK and SEFG planned the experiment. LLV, MSL and SW ran the analyses. LLV, MSL, SW, OM, FFS, MAMG and SEFG contributed to drafting the manuscript. PSL and EFK contributed to the conception of the study and provided data to the writing of the paper. All authors read and approved the final version.

Acknowledgements

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), National Institute of Science and Technology – Animal Science, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/NUFFIC).

References

1. Rodriguez C, Tomas A, Alves E, Ramirez O, Arque M, Munoz G, et al.: **QTL mapping for teat number in an Iberian x Meishan pig intercross**. *Anim Genet* 2005, **36**:490-496.

2. Hirooka H, de Koning DJ, Harlizius B, van Arendonk JA, Rattink AP, Groenen MA, et al.: **A whole-genome scan for quantitative trait loci affecting teat number in pigs**. *J Anim Sci* 2001, **79**(9):2320-2326.

3. Hens JR, Wysolmerski JJ: **Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland**. *Breast Cancer Res* 2005, **7**(5):220.

4. Veltmaat JM, Van Veelen W, Thiery JP, Bellusci S: **Identification of the mammary line in mouse by Wnt10b expression**. *Dev Dynam* 2004, **229**:349-356.

5. Lee MY, Racine V, Jagadpramana P, Sun L, Yu W, Du T, et al.: **Ectodermal Influx and Cell Hypertrophy Provide Early Growth for All Murine Mammary Rudiments, and Are Differentially Regulated among Them by Gli3**. *PLoS One* 2011, **6**:e26242.

6. Christ B, Wilting J: **From somites to vertebral column**. *Ann Anat* 1992, **174**(1):23-32.

7. Ren DR, Ren J, Ruan GF, Guo YM, Wu LH, Yang GC, et al.: **Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc × Chinese Erhualian intercross resource population**. *Anim Genet* 2012, **43**(5):545-551.

8. Lee SS, Chen Y, Moran C, Cepica S, Reiner G, Bartenschlager H, et al.: **Linkage and QTL mapping for *Sus scrofa* chromosome 2**. *J Anim Breed Genet* 2003, **120**(s1):11-19.

9. Guo Y-M, Lee GJ, Archibald AL, Haley CS: **Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan x Large White populations**. *Anim Genet* 2008, **39**(5):486-495.

10. Duijvesteijn N, Veltmaat JM, Knol EF, Harlizius B: **High-resolution association mapping of number of teats in pigs reveals regions controlling vertebral development**. *BMC genomics* 2014, **15**(1):542.

11. Lopes MS, Bastiaansen JW, Harlizius B, Knol EF, Bovenhuis H: **A Genome-Wide Association Study Reveals Dominance Effects on Number of Teats in Pigs**. *PloS one*, **9**(8):e105867.

12. Fortes MR, Reverter-Gomez T, Hiriyur-Nagaraj S, Zhang Y, Jonsson N, Barris W, et al.:**A SNP-derived regulatory gene network underlying puberty in two tropical breeds of beef cattle**. *J Anim Sci* 2011, **89**:1669-1683.

13. Reverter A, Fortes MRS: **Building single nucleotide polymorphism-derived gene regulatory networks: Towards functional genomewide association studies**. *J Anim Sci* 2013, **91**:530-536.

14. Verardo LL, Silva FF, Varona L, Resende MDV, Bastiaansen JWM, Lopes PS, et al.: **Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs**. J App Genet 2014, 1-10.

15. Yu TP, Tuggle CK, Schmitz CB, Rothschild MF: **Association of PIT1 polymorphisms with growth and carcass traits in pigs**. *J Anim Sci* 1995, **73**(5):1282-1288.

16. Chen J, Yang XJ, Xia D, Wegner J, Jiang Z, Zhao RQ: **Sterol regulatory element binding transcription factor 1 expression and genetic polymorphism significantly affect intramuscular fat deposition in the longissimus muscle of Erhualian and Sutai pigs**. *J Anim Scie* 2008, **86**(1):57-63.

17. Veroneze R, Bastiaansen J, Knol EF, Guimarães S, Silva FF, Harlizius B, et al.: **Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations**. *BMC Genet* 2014, **15**(1):126.

18. Machado MA, Azevedo AL, Teodoro RL, Pires MA, Peixoto MG, de Freitas C, et al.: **Genome wide scan for quantitative trait loci affecting tick resistance in cattle (*Bos taurus* x *Bos indicus*)**. *BMC Genomics* 2010, **11**:280.

19. Silva FF, Mulder HA, Knol EF, Lopes MS, Guimarães SE, Lopes PS, et al.: **Sire evaluation for total number born in pigs using a genomic reaction norms approach**. *J Anim Sci* 2014, **92**(9):3825-3834.12.

20. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE et al.: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology**. PLoS ONE 2009, **4**:e6524.

21. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF et al.: **Analyses of pig genomes provide insight into porcine demography and evolution**. *Nature* 2012, **491**(7424):393-398.

22. Mulder HA, Lidauer M, Strandén I, Mäntysaari EA, Pool MH, Veerkamp RF: *MiXBLUP Manual*. Lelystad, the Netherlands: Animal Breeding and Genomics Centre, Wageningen UR Livestock Research; 2012.

23. Tier B, Meyer K: **Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models.** *J Anim Breed Genet* 2004, **121:**77-89.

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

25. George EI, McCulloch RE: **Variable selection via Gibbs sampling.** *J Am Stat Assoc* 1993, **88:**881-889.

26. Gelman A, Rubin DB: **Inference from iterative simulation using multiple sequences.** *Stat Sci* 1992, **7:**457-472.

27. Plummer M, Best N, Cowles K, Vines K: **CODA: Convergence diagnosis and output analysis for MCMC.** *R news* 2006, **6:**7-11.

28. Kass RE, Raftery AE: **Bayes factors.** *J Am Stat Assoc* 1995, 773-795.

29. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**:263-265.

30. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al.: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**:2225-9.

31. Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models**. *Genetics* 1964, **49**:49-67.

32. Bindea G, Mlecnik B, Hack H, Charoentong P, Tosolini M, Kirilovsky A, et al.: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics* 2009, **25(**8):1091-1093.

33. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, (32 Database):D91-4.

34. Touzet H, Varré JS: **Efficient and accurate P-value computation for Position Weight Matrices**. *Algorithms Mol Biol* 2007, **2**(1510.1186):1748-7188.

35. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al.: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genom Res* 2003, **13**(11):2498-2504.

36. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448-3449.

37. Holl JW, Cassady JP, Pomp D, Johnson RK: **A genome scan for quantitative trait loci and imprinted regions affecting reproduction in pigs**. *J Anim Sci* 2004, **82**(12):3421-3429.

38. Bidanel JP, Rosendo A, Iannuccelli N, Riquet J, Gilbert H, Caritez JC, et al.: **Detection of quantitative trait loci for teat number and female reproductive traits in Meishan X Large White F2 pigs**. *Animal* 2008, **2**(6):813.

39. Hernandez SC, Finlayson HA, Ashworth CJ, Haley CS, Archibald AL: **A genome-wide linkage analysis for reproductive traits in F2 Large White× Meishan cross gilts**. *Anim Genet* 2014, **45**(2):191-197.

40. Beeckmann P, Moser G, Bartenschlager H, Reiner G, Geldermann H: **Linkage and QTL mapping for Sus scrofa chromosome 8**. *J Anim Breed Genet* 2003, **120**(s1):66-73.

41. Naureckiene S, Sleat DE, Lackland H, Fensom A, Vanier MT, Wattiaux R, et al: **Identification of HE1 as the second gene of Niemann-Pick C disease**. *Science* 2000, **290**:2298–2301.

42. Vanier MT, Millat G: **Structure and function of the NPC2 protein**. *Biochim Biophys Acta* 2004, **1685**:14–21.

43. Gibson MA, Hatzinikolas G, Davis EC, Baker E, Sutherland GR, Mecham RP: **Bovine latent transforming growth factor beta 1-binding protein 2: molecular cloning, identification of tissue isoforms, and immunolocalization to elastin-associated microfibrils.** *Mol Cell Biol* 1995, **15**:6932–6942.

44. Brancaccio D, Gallo A, Mikolajczyk M, Zovo K, Palumaa P, Novellino E, et al.: **Formation of [4Fe-4S] clusters in the mitochondrial iron-sulfur cluster assembly machinery**. *J Am Chem Soc* 2014, **136**(46):16240–16250.

45. Hebbard LW, Garlatti M, Young LJ, Cardiff RD, Oshima RG, Ranscht B: **T-cadherin supports angiogenesis and adiponectin association with the vasculature in a mouse mammary tumor model**. *Cancer Res* 2008, **68**(5):1407-1416.

46. Hattori K, Heissig B, Wu Y, Dias S, Tejada R, Ferris B, et al.: **Placental growth factor reconstitutes hematopoiesis by recruiting VEGFR1+ stem cells from bone-marrow microenvironment**. *Nature medicine* 2002, **8**(8):841-849.

47. Kuijper S, Turner CJ, Adams RH: **Regulation of angiogenesis by Eph–ephrin interactions**. *Trends in cardiovascular medicine* 2007, **17**(5):145-151.

48. Maloul A, Rossmeier K, Mikic B, Pogue V, Battaglia T: **Geometric and material contributions to whole bone structural behavior in GDF-7-deficient mice**. *Connect Tissue Res* 2006, **47**(3):157-162.

49. Bastian H, Gruss P: **A murine even-skipped homologue, Evx 1, is expressed during early embryogenesis and neurogenesis in a biphasic manner**. *EMBO J* 1990, **9**(6):1839.

50. Tanos T, Rojo LJ, Echeverria P, Brisken C: **ER and PR signaling nodes during mammary gland development**. *Breast Cancer Res* 2012, **14**(4):210.

51. Smith N, Dong Y, Lian JB, Pratap J, Kingsley PD, Van Wijnen AJ, et al.: **Overlapping expression of Runx1 (Cbfa2) and Runx2 (Cbfa1) transcription factors supports cooperative induction of skeletal development**. *J Cell Physiol* 2005, **203**(1):133-143.

52. Wall EH, Case LK, Hewitt SC, Nguyen-Vu T, Candelaria NR, Teuscher C, et al.. **Genetic control of ductal morphology, estrogen-induced ductal growth, and gene expression in female mouse mammary gland**. *Endocrinology* 2014, **155**(8):3025-3035.

53. Miller KA, Eklund EA, Peddinghaus ML, Cao Z, Fernandes N, Turk PW, et al.: **Kruppel-like factor 4 regulates laminin α3a expression in mammary epithelial cells**. *J Biol Chem* 2001, **276**(46):42863-42868.

54. Friedrichs N, Steiner S, Buettner R, Knoepfle G: **Immunohistochemical expression patterns of AP2α and AP2γ in the developing fetal human breast.** *Histopathology* 2007, **51**(6):814-823.

55. Maier JA, Lo Y, Harfe BD: **Foxa1 and Foxa2 are required for formation of the intervertebral discs**. *PloS one* 2013, **8**(1):e55528.

56. Scheijen B, Bronk M, van der Meer T, Bernards R: **Constitutive E2F1 overexpression delays endochondral bone formation by inhibiting chondrocyte differentiation**. *Mol Cell Biol* 2003, **23**(10):3656-3668.

57. El-Magd MA, Allen S, McGonnell I, Mansour AA, Otto A, Patel K: **Shh regulates chick Ebf1 gene expression in somite development**. *Gene* 2014, 554:87–95.

58. Choi YS, Chakrabarti R, Escamilla-Hernandez R, Sinha S: **Elf5 conditional knockout mice reveal its role as a master regulator in mammary alveolar development: failure of Stat5 activation and functional differentiation in the absence of Elf5**. *Dev Biol* 2009, **329**(2):227-241.

59. Truong AB: *Control of epidermal proliferation and differentiation by p63*. Stanford University; 2009.

60. Viganò MA, Lamartine J, Testoni B, Merico D, Alotto D, Castagnoli C, et al.: **New p63 targets in keratinocytes identified by a genome-wide approach**. *EMBO J* 2006, **25**(21):5105-5116.

61. Carroll DK, Carroll JS, Leong CO, Cheng F, Brown M, Mills AA, et al.: **p63 regulates an adhesion programme and cell survival in epithelial cells**. *Nature Cell Biol* 2006, **8**(6):551-561.

62. Dhaka A1, Earley TJ, Watson J, Patapoutian A: **Visualizing cold spots: TRPM8-expressing sensory neurons and their projections**. *J Neurosci* 2008, **28**(3):566-575.

63. Putiri E, Pelegri F: **The zebrafish maternal-effect gene mission impossible encodes the DEAH-box helicase Dhx16 and is essential for the expression of downstream endodermal genes.** *Dev Biol* 2011, **353**(2):275-289.

64. Buchan JG, Gray RS, Gansner JM, Alvarado DM, Burgert L, Gitlin JD, et al.: **Kinesin family member 6 (kif6) is necessary for spine development in zebrafish**. *Dev Dyn* 2014, **243**(12):1646-1657.

65. Miró X, Zhou X, Boretius S, Michaelis T, Kubisch C, Alvarez-Bolado G, et al.: **Haploinsufficiency of the murine polycomb gene Suz12 results in diverse malformations of the brain and neural tube**. *Dis Model Mech* 2009, **2**(7-8):412-418.

66. Vouyovitch CM, Vidal L, Borges S, Raccurt M, Arnould C, Chiesa J, et al.: **Proteomic analysis of autocrine/paracrine effects of human growth hormone in human mammary carcinoma cells**. *Adv Exp Med Biol* 2008, **617**:493-500.

67. Wang Y, Dong J, Li D, Lai L, Siwko S, Li Y, et al.: **Lgr4 regulates mammary gland development and stem cell activity through the pluripotency transcription factor Sox2**. *Stem Cells* 2013, **31**(9):1921-31

68. Topczewska JM, Topczewski J, Shostak A, Kume T, Solnica-Krezel L, Hogan BL: **The winged helix transcription factor Foxc1a is essential for somitogenesis in zebrafish**. *Genes Dev* 2001, **15**(18):2483-2493.

69. Borenstein E, Krakauer DC: **An end to endless forms: Epistasis, phenotype distribution bias, and nonuniform evolution**. *PLoS Comput Biol* 2008, **4**(10): e10000202.

70. Munteanu A, Solé RV: **Neutrality and robustness in Evo-Devo: Emergence of lateral inhibition**. *PLoS Comput Biol* 2008, **4**(11):e10000226.

Additional files

Additional file 1 - Table S1 Pathway and Gene Ontology terms (GO) of genes linked with significant SNPs identified for number of teats in line C

Gene	Pathway	Biological Process (GO)	Molecular Function (GO)	Cellular Component (GO)
CCL20	Signaling by GPCR	chemotaxis/response to biotic stimulus/positive regulation of macromolecule biosynthetic process	receptor binding	extracellular space
CDH13	Wnt signaling pathway	chemotaxis/regulation of endothelial cell proliferation/mitotic cell cycle/positive regulation of macromolecule biosynthetic process	protein homodimerization activity/small GTPase mediated signal transduction	membrane raft/extracellular space
DHX16	Spliceosome	ribonucleotide catabolic process/organonitrogen compound catabolic process/mRNA metabolic process	poly(A) RNA binding/ATP binding	-
EVX1	-	ventral spinal cord development/positive regulation of macromolecule biosynthetic process	dorsal-ventral pattern formation/sequence-specific DNA binding	-
FLOT1	Insulin Signaling	chemotaxis/small GTPase mediated signal transduction	receptor binding/enzyme binding	vacuole/membrane raft
GDF7	Hippo signaling pathway	chemotaxis/gland morphogenesis/positive regulation of macromolecule biosynthetic process/dorsal spinal cord development/protein phosphorylation	receptor binding/protein homodimerization activity	extracellular space
HS1BP3	-	cell activation	lipid binding	-
ILVBL	2-Hydroxyglutric Aciduria (D And L Form)	-	sulfur compound binding	-
ISCA2	-	iron-sulfur cluster assembly	structural molecule activity	-
LOC100523537/ OR1I1	Signaling by GPCR	G-protein coupled receptor signaling pathway	G-protein coupled receptor activity	plasma membrane
LOC100525329/LOC100525507/OR10H3	Signaling by GPCR	neurological system process/detection of chemical stimulus involved in sensory perception of smell	transmembrane signaling receptor activity	-
LOC100525684/LOC100525859/OR10H4	Signaling by GPCR	detection of chemical stimulus involved in sensory perception of smell/neurological system process	transmembrane signaling receptor activity	-

Table S1. Continue

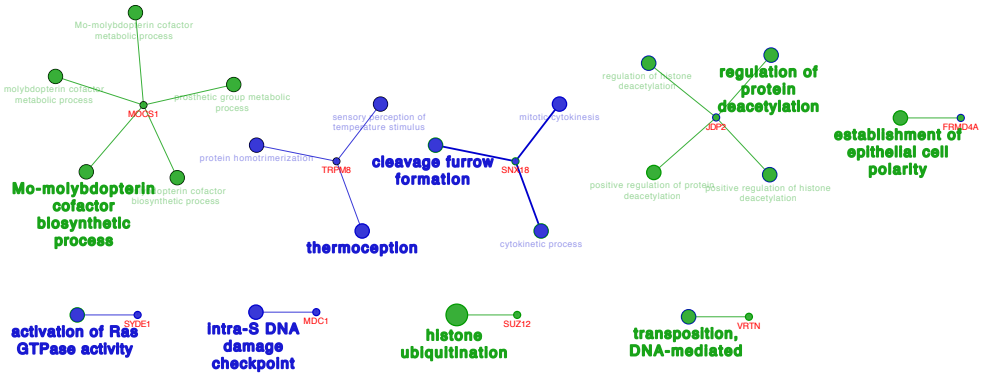
LOC100624137	-	-	-	-
LOC100628051/ OR10T2	Signaling by GPCR	detection of chemical stimulus involved in sensory perception of smell/neurological system process	transmembrane signaling receptor activity	-
LOC100738208/S LCO5A1	-	-	transporter activity	-
LOC100738252	-	-	-	-
LOC102159651/ WBP11	Spliceosome	mRNA metabolic process	poly(A) RNA binding/protein domain specific binding/enzyme regulator activity	nucleoplasm
LOC102165402	-	-	-	-
LTBP2	Extracellular matrix organization	intracellular protein transport	sulfur compound binding	extracellular space
MDC1	DNA Repair	DNA metabolic process/mitotic cell cycle	protein domain specific binding	nucleoplasm/cell junction
NPC2	Lysosome	response to biotic stimulus/cellular lipid metabolic process	lipid binding/enzyme binding	vacuole
NRM	-	-	-	organelle envelope
PPP1R18	-	-	enzyme binding	-
SNX18	-	organonitrogen compound catabolic process/mitotic cell cycle/ribonucleotide catabolic process/intracellular protein transport	lipid binding	cytoplasmic vesicle part
SPP2	-	protein complex biogenesis	enzyme regulator activity	
SYDE1	Rho GTPase cycle	chromosome organization/ribonucleotide catabolic process/organonitrogen compound catabolic process/small GTPase mediated signal transduction	protein complex binding/enzyme regulator activity	synapse
TRPM8	Transmembrane transport of small molecules	neurological system process/protein complex biogenesis/transmembrane transport	transporter activity/protein homodimerization activity	endoplasmic reticulum membrane/membrane raft
TUBB2A	Metabolism of proteins	ribonucleotide catabolic process/microtubule-based process/organonitrogen compound catabolic process	structural molecule activity/guanyl nucleotide binding	microtubule

Additional file 2 - Table S2 Pathway and Gene Ontology terms (GO) of genes linked with significant SNPs identified for number of teats in line D

Gene	Pathway	Biological Process (GO)	Molecular Function (GO)	Cellular Component (GO)
ARFIP1	-	intracellular protein transport	protein domain specific binding	Golgi apparatus
CRLF3	-	positive regulation of macromolecule biosynthetic process/mitotic cell cycle	-	-
EFNB2	Angiogenesis	chemotaxis	receptor binding	integral component of plasma membrane
FRMD4A	-	establishment or maintenance of cell polarity	binding, bridging	cell junction
ISCA2	-	iron-sulfur cluster assembly	structural molecule activity	-
JDP2	Apoptosis signaling pathway	chromosome organization	sequence-specific DNA binding	-
KIF6	-	microtubule-based process	ATP binding/protein complex binding	microtubule
KSR1	Ras signaling pathway	protein phosphorylation/small GTPase mediated signal transduction	structural molecule activity/binding, bridging/ATP binding/enzyme binding	endoplasmic reticulum membrane
LOC100624918/FLVCR2	-	transmembrane transport	transporter activity	-
LOC102165281	-	-	-	-
LOC102165411	-	-	-	-
LOC102165942	-	-	-	-
LTBP2	Extracellular matrix organization	intracellular protein transport	sulfur compound binding	extracellular space
MOCS1	Defective GIF causes intrinsic factor deficiency	coenzyme metabolic process	guanyl nucleotide binding	molybdopterin synthase complex

Table S2. Continue

NPC2	Lysosome	response to biotic stimulus/cellular lipid metabolic process	lipid binding/enzyme binding	vacuole
PGF	Ras signaling pathway	regulation of endothelial cell proliferation	receptor binding/sulfur compound binding/protein homodimerization activity	extracellular space
PGR	Gene Expression	positive regulation of macromolecule biosynthetic process/gland morphogenesis/mammary gland epithelium development	lipid binding/sequence-specific DNA binding/zinc ion binding/receptor binding/enzyme binding	organelle envelope/mitochondrial part/nucleoplasm
PTGR2	-	cellular lipid metabolic process	zinc ion binding	-
RASL11B	-	organonitrogen compound catabolic process/ribonucleotide catabolic process/small GTPase mediated signal transduction	guanyl nucleotide binding	-
RPS6KL1	-	protein phosphorylation	ATP binding	ribonucleoprotein complex
SUZ12	Cellular responses to stress	chromosome organization	sequence-specific DNA binding	nucleoplasm
SYNDIG1L	-	response to biotic stimulus	-	Golgi apparatus
VRTN	-	DNA metabolic process	sequence-specific DNA binding	-



Additional file 3 - Figure S1.pdf
Functionally group networks with biological process terms and genes (labelled in red) as nodes. Blue nodes represents the line C while green nodes the line D. The node size represents the term enrichment significance. The most significant term per group is shown in bold

Additional file 4 - Table S3. Transcription factors (TF) identified for number of teats in C line, their linked genes, the location in base pairs (bp) of the window sequence and the p-value of the window

TF	Genes*	Location of the window (bp)		p-value
RUNX1		3164:2066		0,0000000000003
	LOC100525329/OR10H3-like	2694	2683	
	LOC100525329/OR10H3-like	2362	2351	
	LOC100525507/OR10H3-like	2965	2954	
	LOC100525507/OR10H3-like	2938	2927	
	LOC100525507/OR10H3-like	2913	2902	
	LOC100525507/OR10H3-like	2457	2446	
	LOC100525684/OR10H4-like	2911	2900	
	LOC100525684/OR10H4-like	2648	2637	
	LOC100525684/OR10H4-like	2084	2073	
	LOC100525859/OR10H4-like	2478	2467	
	LOC100525859/OR10H4-like	2094	2083	
	LOC100628051/OR10T2-like	2344	2333	
	LOC100628051/OR10T2-like	2173	2162	
	LOC100628051/OR10T2-like	2105	2094	
	ILVBL	3164	3153	
	ILVBL	2840	2829	
	ILVBL	2794	2783	
	ILVBL	2582	2571	
	ILVBL	2274	2263	
	SYDE1	2550	2539	
	SYDE1	2079	2068	
	LOC100523537/OR1I1-like	3041	3030	
	LOC100523537/OR1I1-like	2846	2835	
	LOC100523537/OR1I1-like	2754	2743	
	LOC100523537/OR1I1-like	2536	2525	

LOC100523537/OR1I1-like	2334	2323
LOC100523537/OR1I1-like	2172	2161
GDF7	3152	3141
GDF7	2528	2517
LOC102159651/WBP11-like	2278	2267
LOC102159651/WBP11-like	2245	2234
HS1BP3	2107	2096
LOC100738208/SLCO5A1-like	3154	3143
LOC100738208/SLCO5A1-like	3114	3103
LOC100738208/SLCO5A1-like	2787	2776
LOC100738208/SLCO5A1-like	2690	2679
LOC100738208/SLCO5A1-like	2656	2645
LOC100738208/SLCO5A1-like	2318	2307
LOC100738252	3136	3125
LOC100738252	3075	3064
LOC100738252	2685	2674
LOC100738252	2112	2101
LOC100738252	2066	2055
SLCO5A1	2267	2256
SLCO5A1	2227	2216
SLCO5A1	2136	2125
CDH13	2554	2543
DHX16	3156	3145
DHX16	3063	3052
DHX16	2737	2726
DHX16	2680	2669
DHX16	2640	2629
DHX16	2417	2406
DHX16	2403	2392
DHX16	2162	2151
DHX16	2102	2091
PPP1R18	2781	2770
PPP1R18	2584	2573
PPP1R18	2352	2341
PPP1R18	2167	2156
PPP1R18	2071	2060
NRM	3053	3042
NRM	3006	2995
NRM	2844	2833
NRM	2753	2742
NRM	2713	2702
NRM	2377	2366
NRM	2289	2278
MDC1	2958	2947
MDC1	2824	2813
MDC1	2710	2699
MDC1	2223	2212
TUBB2A	2161	2150
FLOT1	2936	2925
FLOT1	2592	2581

	FLOT1	2373	2362	
	FLOT1	2344	2333	
	FLOT1	2260	2249	
	FLOT1	2235	2224	
	NPC2	2666	2655	
	NPC2	2209	2198	
	ISCA2	3108	3097	
	ISCA2	3062	3051	
	ISCA2	3016	3005	
	ISCA2	2864	2853	
	ISCA2	2699	2688	
	ISCA2	2442	2431	
	ISCA2	2129	2118	
	ISCA2	2077	2066	
	LTBP2	2981	2970	
	LTBP2	2723	2712	
	LTBP2	2546	2535	
	LTBP2	2117	2106	
	LOC102165402	3130	3119	
	LOC102165402	3000	2989	
	LOC102165402	2960	2949	
	LOC102165402	2903	2892	
	LOC102165402	2862	2851	
	CCL20	2423	2412	
	CCL20	2383	2372	
	TRPM8	3126	3115	
	TRPM8	2990	2979	
	TRPM8	2884	2873	
	TRPM8	2662	2651	
	TRPM8	2571	2560	
	TRPM8	2531	2520	
	TRPM8	2254	2243	
	TRPM8	2106	2095	
	SPP2	3136	3125	
	SPP2	3009	2998	
	SPP2	2961	2950	
	SPP2	2066	2055	
	SNX18	2808	2797	
	SNX18	2557	2546	
KLF4		430:329		0,0000003
	LOC100628051/OR10T2-like	397	387	
	ILVBL	418	408	
	SYDE1	329	319	
	GDF7	430	420	
	GDF7	378	368	
	GDF7	355	345	
	LOC102159651/WBP11-like	404	394	
	HS1BP3	367	357	
	HS1BP3	351	341	
	HS1BP3	330	320	

	LOC100738208/SLCO5A1-like	404	394	
	LOC100738208/SLCO5A1-like	381	371	
	LOC100738208/SLCO5A1-like	355	345	
	LOC100738208/SLCO5A1-like	342	332	
	SLCO5A1	404	394	
	PPP1R18	406	396	
	NRM	405	395	
	NRM	393	383	
	NRM	363	353	
	NRM	333	323	
	FLOT1	430	420	
	FLOT1	418	408	
	FLOT1	406	396	
	FLOT1	360	350	
	NPC2	415	405	
	NPC2	358	348	
	NPC2	347	337	
	ISCA2	426	416	
	ISCA2	343	333	
	LTBP2	429	419	
	LTBP2	365	355	
	CCL20	390	380	
	TRPM8	331	321	
	SPP2	370	360	
	SNX18	335	325	
ZNF354C		139:15		0,000004
	LOC100525329/OR10H3-like	15	9	
	LOC100525684/OR10H4-like	135	129	
	LOC100525684/OR10H4-like	100	94	
	LOC100525684/OR10H4-like	40	34	
	SYDE1	139	133	
	LOC100523537/OR1I1-like	138	132	
	HS1BP3	47	41	
	SLCO5A1	19	13	
	NRM	127	121	
	NRM	26	20	
	MDC1	89	83	
	MDC1	67	61	
	ISCA2	107	101	
	ISCA2	71	65	
	TRPM8	52	46	
RUNX1		908:832		0,000004
	LOC100525329/OR10H3-like	833	822	
	LOC100525684/OR10H4-like	888	877	
	ILVBL	874	863	
	ILVBL	834	823	
	SYDE1	904	893	
	SYDE1	853	842	
	GDF7	848	837	
	LOC102159651/WBP11-like	908	897	

	LOC102159651/WBP11-like	874	863	
	LOC102159651/WBP11-like	859	848	
	HS1BP3	832	821	
	SLCO5A1	860	849	
	DHX16	850	839	
	NRM	839	828	
	ISCA2	863	852	
SP1		407:324		0,000005
	LOC100628051/OR10T2-like	396	386	
	ILVBL	334	324	
	SYDE1	407	397	
	SYDE1	334	324	
	GDF7	379	369	
	GDF7	369	359	
	GDF7	359	349	
	LOC102159651/WBP11-like	403	393	
	HS1BP3	373	363	
	HS1BP3	352	342	
	HS1BP3	331	321	
	LOC100738208/SLCO5A1-like	380	370	
	LOC100738208/SLCO5A1-like	348	338	
	CDH13	403	393	
	DHX16	376	366	
	PPP1R18	407	397	
	PPP1R18	397	387	
	PPP1R18	359	349	
	NRM	404	394	
	NRM	364	354	
	NRM	327	317	
	FLOT1	407	397	
	FLOT1	361	351	
	NPC2	357	347	
	ISCA2	344	334	
	ISCA2	324	314	
	LTBP2	400	390	
	LTBP2	342	332	
	LTBP2	325	315	
	TRPM8	382	372	
	TRPM8	359	349	
	TRPM8	330	320	
	SPP2	372	362	
	SNX18	334	324	
	EVX1	359	349	
	EVX1	326	316	
FOXA2		2597:2369		0,000007
	LOC100525507/OR10H3-like	2403	2391	
	LOC100628051/OR10T2-like	2476	2464	
	ILVBL	2590	2578	
	ILVBL	2547	2535	
	ILVBL	2522	2510	

	ILVBL	2462	2450	
	SYDE1	2515	2503	
	LOC100738208/SLCO5A1-like	2500	2488	
	LOC100738252	2474	2462	
	SLCO5A1	2592	2580	
	SLCO5A1	2491	2479	
	CDH13	2548	2536	
	CDH13	2420	2408	
	DHX16	2537	2525	
	PPP1R18	2565	2553	
	PPP1R18	2437	2425	
	NRM	2597	2585	
	NRM	2585	2573	
	MDC1	2549	2537	
	MDC1	2449	2437	
	FLOT1	2523	2511	
	NPC2	2553	2541	
	NPC2	2386	2374	
	ISCA2	2591	2579	
	ISCA2	2398	2386	
	LTBP2	2376	2364	
	TRPM8	2424	2412	
	TRPM8	2369	2357	
	SNX18	2382	2370	
	EVX1	2483	2471	
GABPA		3089:3040		0,000009
	LOC100628051/OR10T2-like	3089	3078	
	SYDE1	3067	3056	
	CDH13	3075	3064	
	DHX16	3069	3058	
	LTBP2	3041	3030	
	CCL20	3078	3067	
	TRPM8	3062	3051	
	EVX1	3081	3070	
	EVX1	3040	3029	
HNF4A		1588:1239		0,00001
	LOC100525329/OR10H3-like	1374	1361	
	LOC100525507/OR10H3-like	1367	1354	
	LOC100525684/OR10H4-like	1453	1440	
	LOC100628051/OR10T2-like	1528	1515	
	ILVBL	1537	1524	
	SYDE1	1325	1312	
	LOC100523537/OR1I1-like	1275	1262	
	HS1BP3	1495	1482	
	HS1BP3	1268	1255	
	LOC100738208/SLCO5A1-like	1587	1574	
	LOC100738252	1364	1351	
	SLCO5A1	1525	1512	
	SLCO5A1	1398	1385	
	CDH13	1297	1284	

	CDH13	1240	1227	
	DHX16	1580	1567	
	DHX16	1476	1463	
	DHX16	1426	1413	
	DHX16	1350	1337	
	PPP1R18	1284	1271	
	NRM	1542	1529	
	NRM	1318	1305	
	FLOT1	1439	1426	
	NPC2	1572	1559	
	ISCA2	1239	1226	
	LTBP2	1274	1261	
	LOC102165402	1497	1484	
	CCL20	1526	1513	
	TRPM8	1457	1444	
	SPP2	1320	1307	
	SNX18	1588	1575	
REL		2333:2050		0,00001
	LOC100525507/OR10H3-like	2077	2067	
	LOC100525859/OR10H4-like	2333	2323	
	GDF7	2095	2085	
	HS1BP3	2234	2224	
	LOC100738208/SLCO5A1-like	2263	2253	
	LOC100738252	2276	2266	
	SLCO5A1	2321	2311	
	DHX16	2150	2140	
	NRM	2332	2322	
	TUBB2A	2221	2211	
	TUBB2A	2127	2117	
	FLOT1	2184	2174	
	FLOT1	2088	2078	
	ISCA2	2320	2310	
	ISCA2	2079	2069	
	LTBP2	2133	2123	
	LTBP2	2050	2040	
	CCL20	2327	2317	
	CCL20	2205	2195	
	TRPM8	2328	2318	
	SNX18	2228	2218	
	SNX18	2066	2056	
	EVX1	2326	2316	
MZF1_1		1205:1173		0,00002
	LOC102159651/WBP11-like	1176	1170	
	CDH13	1185	1179	
	TUBB2A	1188	1182	
	FLOT1	1205	1199	
	TRPM8	1191	1185	
	SPP2	1186	1180	
	SNX18	1192	1186	
	EVX1	1173	1167	

NFKB1		2334:2291		0,00002
	LOC100525859/OR10H4-like	2333	2322	
	SLCO5A1	2320	2309	
	NRM	2334	2323	
	MDC1	2291	2280	
	LTBP2	2291	2280	
	CCL20	2329	2318	
	TRPM8	2329	2318	
	EVX1	2327	2316	
SPI1		3190:3059		0,00002
	LOC100525329/OR10H3-like	3070	3063	
	LOC100525507/OR10H3-like	3065	3058	
	LOC100525684/OR10H4-like	3190	3183	
	LOC100628051/OR10T2-like	3188	3181	
	LOC100628051/OR10T2-like	3086	3079	
	SYDE1	3066	3059	
	HS1BP3	3119	3112	
	LOC100738252	3123	3116	
	SLCO5A1	3140	3133	
	CDH13	3072	3065	
	DHX16	3086	3079	
	PPP1R18	3161	3154	
	TUBB2A	3111	3104	
	FLOT1	3184	3177	
	ISCA2	3188	3181	
	LTBP2	3179	3172	
	LOC102165402	3178	3171	
	CCL20	3075	3068	
	TRPM8	3059	3052	
	SPP2	3140	3133	
	EVX1	3080	3073	
E2F1		1037:891		0,00003
	LOC100525329/OR10H3-like	972	964	
	LOC100525507/OR10H3-like	1037	1029	
	ILVBL	1027	1019	
	LOC102159651/WBP11-like	907	899	
	MDC1	983	975	
	NPC2	927	919	
	NPC2	891	883	
	ISCA2	1015	1007	
	SPP2	981	973	
	SNX18	968	960	
NFE2L1::MAFG		605:128		0,00003
	LOC100525329/OR10H3-like	196	190	
	LOC100525507/OR10H3-like	422	416	
	LOC100525507/OR10H3-like	408	402	
	LOC100525684/OR10H4-like	301	295	
	LOC100525684/OR10H4-like	163	157	
	LOC100628051/OR10T2-like	301	295	

	SYDE1	561	555	
	HS1BP3	262	256	
	CDH13	605	599	
	CDH13	330	324	
	CDH13	128	122	
	MDC1	356	350	
	TUBB2A	332	326	
	LTBP2	583	577	
	SPP2	568	562	
	SPP2	338	332	
	EVX1	575	569	
EGR1		571:328		0,00004
	ILVBL	542	531	
	SYDE1	571	560	
	SYDE1	337	326	
	GDF7	498	487	
	GDF7	376	365	
	LOC102159651/WBP11-like	359	348	
	HS1BP3	500	489	
	LOC100738208/SLCO5A1-like	359	348	
	SLCO5A1	464	453	
	SLCO5A1	408	397	
	DHX16	556	545	
	DHX16	522	511	
	DHX16	382	371	
	DHX16	329	318	
	NRM	569	558	
	NRM	421	410	
	MDC1	450	439	
	FLOT1	434	423	
	FLOT1	364	353	
	NPC2	570	559	
	NPC2	546	535	
	NPC2	419	408	
	ISCA2	566	555	
	ISCA2	439	428	
	ISCA2	415	404	
	LTBP2	426	415	
	LTBP2	340	329	
	LOC102165402	520	509	
	TRPM8	328	317	
	SNX18	427	416	
	SNX18	352	341	
PAX6		2268:2230		0,00005
	LOC100525507/OR10H3-like	2239	2225	
	LOC100525684/OR10H4-like	2268	2254	
	GDF7	2268	2254	
	HS1BP3	2266	2252	
	LOC100738252	2260	2246	
	ISCA2	2230	2216	

	CCL20	2231	2217	
LHX3		2616:2372		0,00005
	LOC100525684/OR10H4-like	2616	2603	
	LOC100525684/OR10H4-like	2529	2516	
	LOC100525859/OR10H4-like	2434	2421	
	ILVBL	2407	2394	
	SYDE1	2457	2444	
	SYDE1	2443	2430	
	GDF7	2493	2480	
	LOC100738208/SLCO5A1-like	2414	2401	
	LOC100738252	2412	2399	
	SLCO5A1	2539	2526	
	CDH13	2603	2590	
	CDH13	2528	2515	
	DHX16	2543	2530	
	NRM	2609	2596	
	NRM	2595	2582	
	NRM	2581	2568	
	MDC1	2446	2433	
	FLOT1	2535	2522	
	FLOT1	2497	2484	
	NPC2	2608	2595	
	NPC2	2419	2406	
	NPC2	2384	2371	
	ISCA2	2592	2579	
	LOC102165402	2375	2362	
	CCL20	2569	2556	
	TRPM8	2405	2392	
	TRPM8	2372	2359	
SRY		925:744		0,00006
	LOC100525329/OR10H3-like	925	916	
	LOC100525507/OR10H3-like	784	775	
	LOC100525859/OR10H4-like	906	897	
	LOC100525859/OR10H4-like	828	819	
	SYDE1	878	869	
	SYDE1	793	784	
	LOC100523537/OR1I1-like	820	811	
	LOC100523537/OR1I1-like	769	760	
	HS1BP3	838	829	
	LOC100738208/SLCO5A1-like	868	859	
	LOC100738208/SLCO5A1-like	858	849	
	LOC100738252	745	736	
	SLCO5A1	773	764	
	CDH13	800	791	
	MDC1	744	735	
	TUBB2A	882	873	
	TUBB2A	849	840	
	TUBB2A	840	831	
	CCL20	826	817	
	SPP2	871	862	

	SPP2	759	750	
	EVX1	925	916	
	EVX1	892	883	
NR1H2::RXRA		2517:2232		0,00006
	LOC100525329/OR10H3-like	2260	2243	
	ILVBL	2344	2327	
	ILVBL	2285	2268	
	SYDE1	2264	2247	
	LOC100523537/OR1I1-like	2479	2462	
	GDF7	2510	2493	
	GDF7	2310	2293	
	LOC102159651/WBP11-like	2282	2265	
	HS1BP3	2516	2499	
	HS1BP3	2466	2449	
	HS1BP3	2323	2306	
	HS1BP3	2289	2272	
	LOC100738208/SLCO5A1-like	2351	2334	
	SLCO5A1	2517	2500	
	MDC1	2455	2438	
	NPC2	2236	2219	
	ISCA2	2232	2215	
	LTBP2	2444	2427	
	LOC102165402	2295	2278	
	TRPM8	2493	2476	
	SNX18	2500	2483	
	EVX1	2374	2357	
NFKB1		3188:3056		0,00007
	SYDE1	3161	3150	
	LOC100738208/SLCO5A1-like	3061	3050	
	DHX16	3169	3158	
	PPP1R18	3162	3151	
	NRM	3169	3158	
	NRM	3100	3089	
	TUBB2A	3168	3157	
	ISCA2	3159	3148	
	SPP2	3077	3066	
	SPP2	3056	3045	
	EVX1	3188	3177	
	EVX1	3117	3106	
	EVX1	3081	3070	
INSM1		1675:1453		0,00008
	LOC100525684/OR10H4-like	1453	1441	
	LOC100628051/OR10T2-like	1626	1614	
	LOC100628051/OR10T2-like	1585	1573	
	ILVBL	1488	1476	
	LOC102159651/WBP11-like	1490	1478	
	HS1BP3	1484	1472	
	LOC100738208/SLCO5A1-like	1489	1477	
	SLCO5A1	1525	1513	

	DHX16	1476	1464	
	PPP1R18	1675	1663	
	NRM	1648	1636	
	NRM	1622	1610	
	NRM	1541	1529	
	FLOT1	1621	1609	
	FLOT1	1534	1522	
	FLOT1	1486	1474	
	NPC2	1624	1612	
	NPC2	1495	1483	
	LTBP2	1662	1650	
	CCL20	1652	1640	
	TRPM8	1604	1592	
	TRPM8	1456	1444	
	EVX1	1668	1656	
	EVX1	1606	1594	
TFAP2A		1809:1567		0,00009
	LOC100525329/OR10H3-like	1746	1737	
	LOC100628051/OR10T2-like	1699	1690	
	LOC100628051/OR10T2-like	1683	1674	
	LOC100628051/OR10T2-like	1661	1652	
	LOC100628051/OR10T2-like	1641	1632	
	LOC100628051/OR10T2-like	1608	1599	
	LOC100628051/OR10T2-like	1577	1568	
	LOC100628051/OR10T2-like	1567	1558	
	LOC100523537/OR1I1-like	1791	1782	
	HS1BP3	1794	1785	
	PPP1R18	1572	1563	
	MDC1	1809	1800	
	MDC1	1645	1636	
	FLOT1	1792	1783	
	FLOT1	1771	1762	
	FLOT1	1723	1714	
	LOC102165402	1776	1767	
	EVX1	1735	1726	
SRF		466:405		0,00009
	LOC100525329/OR10H3-like	446	434	
	LOC100525507/OR10H3-like	405	393	
	LOC100525859/OR10H4-like	408	396	
	LOC100738252	439	427	
	DHX16	437	425	
	LOC102165402	439	427	
	SNX18	466	454	
E2F1		1526:1388		0,0001
	LOC100525329/OR10H3-like	1479	1471	
	LOC100525684/OR10H4-like	1417	1409	
	LOC100628051/OR10T2-like	1388	1380	
	GDF7	1436	1428	
	CDH13	1526	1518	
	DHX16	1518	1510	

	DHX16	1445	1437	
	TRPM8	1450	1442	
MYCN		1871:1516		0,0001
	GDF7	1558	1548	
	HS1BP3	1612	1602	
	LOC100738208/SLCO5A1-like	1516	1506	
	LOC100738252	1861	1851	
	SLCO5A1	1683	1673	
	PPP1R18	1530	1520	
	MDC1	1767	1757	
	MDC1	1752	1742	
	FLOT1	1811	1801	
	LTBP2	1546	1536	
	CCL20	1871	1861	
	CCL20	1602	1592	
	TRPM8	1779	1769	
	SPP2	1602	1592	
	EVX1	1687	1677	

Additional file 5 - Table S4. Transcription factors (TF) identified for number of teats in D line, their linked genes, the location in base pairs (bp) of the window sequence and the p-value of the window

TF	Gene	Location of the window (bp)		p-value
RUNX1		1662:1016		0.0000003
	KIF6	1662	1651	
	KIF6	1574	1563	
	KIF6	1526	1515	
	KIF6	1469	1458	
	KIF6	1428	1417	
	KIF6	1090	1079	
	KIF6	1070	1059	
	MOCS1	1430	1419	
	MOCS1	1063	1052	
	MOCS1	1016	1005	
	VRTN	1354	1343	
	VRTN	1065	1054	
	VRTN	1034	1023	
	SYNDIG1L	1427	1416	
	SYNDIG1L	1241	1230	
	NPC2	1336	1325	
	ISCA2	1611	1600	
	ISCA2	1218	1207	
	ISCA2	1194	1183	
	LTBP2	1636	1625	
	RPS6KL1	1513	1502	
	RPS6KL1	1229	1218	
	RPS6KL1	1018	1007	
	PGF	1068	1057	

	PGF	1025	1014	
	RASL11B	1565	1554	
	RASL11B	1152	1141	
	ARFIP1	1403	1392	
	ARFIP1	1112	1101	
	PGR	1216	1205	
	FRMD4A	1378	1367	
	EFNB2	1657	1646	
	EFNB2	1559	1548	
	CRLF3	1640	1629	
	CRLF3	1579	1568	
	CRLF3	1545	1534	
	CRLF3	1533	1522	
	CRLF3	1443	1432	
	CRLF3	1192	1181	
	SUZ12	1535	1524	
	SUZ12	1158	1147	
	KSR1	1646	1635	
	KSR1	1606	1595	
	KSR1	1515	1504	
	KSR1	1331	1320	
ETS1		810:368		0.0000009
	KIF6	760	754	
	KIF6	637	631	
	KIF6	517	511	
	KIF6	410	404	
	MOCS1	553	547	
	PTGR2	805	799	
	VRTN	769	763	
	VRTN	740	734	
	VRTN	590	584	
	VRTN	502	496	
	VRTN	395	389	
	SYNDIG1L	810	804	
	SYNDIG1L	593	587	
	SYNDIG1L	419	413	
	SYNDIG1L	394	388	
	NPC2	787	781	
	NPC2	610	604	
	NPC2	519	513	
	NPC2	402	396	
	NPC2	368	362	
	ISCA2	796	790	
	ISCA2	479	473	
	LTBP2	776	770	
	LTBP2	700	694	

	LTBP2	681	675	
	LTBP2	484	478	
	RPS6KL1	795	789	
	RPS6KL1	514	508	
	PGF	447	441	
	JDP2	596	590	
	FLVCR2-like	776	770	
	FLVCR2-like	375	369	
	FLVCR2-like	369	363	
	RASL11B	435	429	
	ARFIP1	754	748	
	ARFIP1	388	382	
	PGR	521	515	
	PGR	396	390	
	FRMD4A	787	781	
	FRMD4A	654	648	
	CRLF3	561	555	
	SUZ12	740	734	
	SUZ12	701	695	
	SUZ12	515	509	
	KSR1	667	661	
	KSR1	650	644	
REST		1374:1268		0.00002
	KIF6	1364	1343	
	VRTN	1269	1248	
	SYNDIG1L	1343	1322	
	SYNDIG1L	1269	1248	
	LTBP2	1332	1311	
	RPS6KL1	1367	1346	
	RASL11B	1268	1247	
	PGR	1364	1343	
	PGR	1284	1263	
	FRMD4A	1374	1353	
	EFNB2	1324	1303	
	SUZ12	1282	1261	
	KSR1	1370	1349	
RELA		1965:1914		0.00002
	KIF6	1940	1930	
	LTBP2	1933	1923	
	FLVCR2-like	1914	1904	
	PGR	1947	1937	
	FRMD4A	1965	1955	
	CRLF3	1938	1928	
	SUZ12	1916	1906	
	KSR1	1961	1951	
MIZF		835:491		0.00003

	SYNDIG1L	835	825	
	SYNDIG1L	657	647	
	SYNDIG1L	494	484	
	NPC2	824	814	
	NPC2	532	522	
	RPS6KL1	491	481	
	PGF	649	639	
	PGF	496	486	
	JDP2	723	713	
	JDP2	622	612	
	FLVCR2-like	500	490	
	EFNB2	611	601	
RREB1		915:697		0.00004
	KIF6	725	705	
	MOCS1	768	748	
	MOCS1	726	706	
	PTGR2	915	895	
	PTGR2	860	840	
	PTGR2	829	809	
	PTGR2	803	783	
	PTGR2	724	704	
	VRTN	870	850	
	LTBP2	912	892	
	RPS6KL1	907	887	
	RPS6KL1	756	736	
	PGF	887	867	
	PGF	867	847	
	PGF	759	739	
	PGF	737	717	
	JDP2	837	817	
	FLVCR2-like	697	677	
	FRMD4A	715	695	
	EFNB2	895	875	
	EFNB2	699	679	
	CRLF3	906	886	
	SUZ12	761	741	
	KSR1	749	729	
ETS1		1694:1505		0.0001
	NPC2	1513	1507	
	RPS6KL1	1641	1635	
	PGF	1694	1688	
	PGF	1614	1608	
	JDP2	1687	1681	
	FRMD4A	1505	1499	
GATA3		1668:1522		0.0001
	MOCS1	1598	1592	

	VRTN	1651	1645	
	ISCA2	1592	1586	
	FLVCR2-like	1668	1662	
	RASL11B	1658	1652	
	RASL11B	1529	1523	
	ARFIP1	1640	1634	
	ARFIP1	1522	1516	
	CRLF3	1588	1582	
TBP		1577:1191		0.0001
	KIF6	1391	1376	
	KIF6	1366	1351	
	PTGR2	1508	1493	
	PTGR2	1352	1337	
	PTGR2	1250	1235	
	SYNDIG1L	1555	1540	
	NPC2	1258	1243	
	NPC2	1214	1199	
	NPC2	1191	1176	
	ISCA2	1332	1317	
	RPS6KL1	1558	1543	
	JDP2	1577	1562	
	JDP2	1358	1343	
	JDP2	1210	1195	
	FLVCR2-like	1448	1433	
	RASL11B	1551	1536	
	PGR	1499	1484	
	PGR	1340	1325	
	PGR	1212	1197	
	SUZ12	1479	1464	
	SUZ12	1333	1318	
EBF1		703:425		0.0002
	MOCS1	681	671	
	PTGR2	469	459	
	VRTN	489	479	
	NPC2	656	646	
	ISCA2	581	571	
	ISCA2	443	433	
	RPS6KL1	687	677	
	PGF	703	693	
	PGF	510	500	
	JDP2	498	488	
	JDP2	439	429	
	RASL11B	650	640	
	RASL11B	632	622	
	RASL11B	618	608	
	RASL11B	549	539	

	ARFIP1	529	519	
	PGR	435	425	
	FRMD4A	687	677	
	EFNB2	498	488	
	SUZ12	425	415	
	KSR1	470	460	
Pax6		66:35*		0.0002
	RPS6KL1	35	49*	
	RASL11B	66	52	
	ARFIP1	14	0	
	FRMD4A	57	43	
	FRMD4A	6	0	
	EFNB2	47	33	
	KSR1	5	19*	
REL		1281:948		0.0002
	PTGR2	1253	1243	
	PTGR2	1206	1196	
	PTGR2	1000	990	
	VRTN	1280	1270	
	VRTN	1049	1039	
	NPC2	1281	1271	
	NPC2	1038	1028	
	NPC2	960	950	
	ISCA2	1241	1231	
	PGF	1205	1195	
	FLVCR2-like	1252	1242	
	RASL11B	1060	1050	
	ARFIP1	1016	1006	
	ARFIP1	948	938	
	PGR	1119	1109	
	FRMD4A	1155	1145	
	EFNB2	1209	1199	
	EFNB2	1197	1187	
	CRLF3	950	940	
FOXF2		670:496		0.0002
	KIF6	499	485	
	MOCS1	631	617	
	MOCS1	506	492	
	RPS6KL1	540	526	
	PGF	534	520	
	JDP2	663	649	
	JDP2	514	500	
	FLVCR2-like	589	575	
	ARFIP1	670	656	
	ARFIP1	620	606	
	ARFIP1	588	574	

	ARFIP1	553	539	
	ARFIP1	496	482	
	PGR	631	617	
MZF1_5-13		1082:1014		0.0003
	KIF6	1014	1004	
	VRTN	1057	1047	
	VRTN	1017	1007	
	SYNDIG1L	1056	1046	
	FLVCR2-like	1082	1072	
	FRMD4A	1061	1051	
	EFNB2	1067	1057	
	EFNB2	1041	1031	
	KSR1	1077	1067	
MZF1_5-13		1732:1468		0.0003
	KIF6	1488	1478	
	MOCS1	1679	1669	
	PTGR2	1596	1586	
	SYNDIG1L	1728	1718	
	SYNDIG1L	1578	1568	
	NPC2	1481	1471	
	LTBP2	1729	1719	
	LTBP2	1669	1659	
	LTBP2	1644	1634	
	LTBP2	1491	1481	
	RPS6KL1	1732	1722	
	RPS6KL1	1535	1525	
	PGF	1588	1578	
	ARFIP1	1480	1470	
	FRMD4A	1654	1644	
	FRMD4A	1630	1620	
	FRMD4A	1541	1531	
	EFNB2	1531	1521	
	EFNB2	1504	1494	
	SUZ12	1468	1458	
GATA3		496:228		0.0004
	VRTN	379	373	
	VRTN	228	222	
	NPC2	259	253	
	ISCA2	321	315	
	LTBP2	444	438	
	PGF	266	260	
	FLVCR2-like	465	459	
	ARFIP1	240	234	
	PGR	292	286	
	FRMD4A	496	490	
Klf4		166:114		0.0004

	KIF6	166	156	
	MOCS1	128	118	
	MOCS1	116	106	
	VRTN	142	132	
	NPC2	158	148	
	NPC2	147	137	
	ISCA2	143	133	
	LTBP2	165	155	
	JDP2	159	149	
	FLVCR2-like	158	148	
	RASL11B	149	139	
	EFNB2	116	106	
	CRLF3	124	114	
	CRLF3	114	104	
	SUZ12	115	105	
Hltf		1716:1639		0.0004
	PTGR2	1659	1649	
	VRTN	1715	1705	
	VRTN	1640	1630	
	SYNDIG1L	1676	1666	
	NPC2	1711	1701	
	JDP2	1654	1644	
	JDP2	1639	1629	
	FLVCR2-like	1653	1643	
	PGR	1716	1706	
	PGR	1692	1682	
	PGR	1647	1637	
Pax2		759:179		0.0005
	PTGR2	574	566	
	NPC2	179	171	
	ISCA2	403	395	
	RPS6KL1	555	547	
	RPS6KL1	239	231	
	PGF	725	717	
	JDP2	759	751	
	JDP2	566	558	
	JDP2	204	196	
	JDP2	188	180	
	FLVCR2-like	255	247	
	ARFIP1	415	407	
	FRMD4A	673	665	
	FRMD4A	524	516	
	CRLF3	642	634	
	CRLF3	537	529	
	SUZ12	194	186	
	KSR1	275	267	

ELF5		1927:1893		0.0005
	KIF6	1916	1907	
	ISCA2	1918	1909	
	ARFIP1	1910	1901	
	PGR	1893	1884	
	SUZ12	1927	1918	
	SUZ12	1913	1904	
CREB1		803:643		0.0005
	MOCS1	803	795	
	VRTN	791	783	
	NPC2	729	721	
	NPC2	693	685	
	ISCA2	660	652	
	LTBP2	710	702	
	RPS6KL1	662	654	
	PGF	656	648	
	KSR1	643	635	
TFAP2A		639:580		0.0006
	MOCS1	591	582	
	PTGR2	639	630	
	VRTN	632	623	
	NPC2	580	571	
	ISCA2	600	591	
	LTBP2	597	588	
	RASL11B	630	621	
	RASL11B	602	593	
	EFNB2	588	579	
	SUZ12	611	602	
STAT1		1036:996		0.0007
	KIF6	1001	986	
	NPC2	1036	1021	
	NPC2	996	981	
	ISCA2	1031	1016	
	JDP2	1009	994	
	CRLF3	1015	1000	
Myc		755:649		0.0007
	MOCS1	657	647	
	VRTN	755	745	
	EFNB2	723	713	
	EFNB2	664	654	
	CRLF3	698	688	
	CRLF3	649	639	
	KSR1	691	681	
RELA		1288:950		0.0007
	MOCS1	1276	1266	
	PTGR2	1253	1243	

PTGR2	1206	1196
PTGR2	1001	991
VRTN	1280	1270
VRTN	1049	1039
SYNDIG1L	952	942
NPC2	1281	1271
NPC2	1038	1028
NPC2	960	950
ISCA2	1076	1066
RASL11B	1060	1050
ARFIP1	1016	1006
PGR	1288	1278
PGR	1119	1109
FRMD4A	1155	1145
EFNB2	1198	1188
CRLF3	950	940

*Downstream location

Chapter 5

Gene networks from candidate genes for total number born in pigs across divergent environments

Lucas L Verardo[1, 2*], Marcos S Lopes[2, 3], Pramod Mathur[3], Ole Madsen[2], Fabyano F Silva[1], Martien AM Groenen[2], Egbert F Knol[3], Paulo S Lopes[1], Simone EF Guimarães[1]
[1]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, 36570000, Brazil
[2] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6700 AH, the Netherlands
[3] Topigs Norsvin Research Center, Beuningen, 6641 SZ, the Netherlands

(To be submitted)

Abstract

A large proportion of the variance on reproductive traits, such as total number born (TNB), are due to environment being highly relevant in pig breeding. Under a random regression reaction norm (RRRN) models approach, it is possible to estimate the additive genomic breeding values for animals over different environmental gradients, and furthermore it is also possible to estimate single nucleotides polymorphism (SNP) effects over such environments. In addition, gene networks from genome wide association studies (GWAS) can be performed using genes related with identified SNP to examine the sharing of pathways and functions involving these genes. The objective of this work was to present gene network analysis for TNB in pigs across different environment levels using results from a random regression reaction norm model GWAS approach. Under this framework, relevant SNPs were identified as their linked genes for each environment separately. Different sets of relevant SNPs and QTL blocks across environments were identified leading to the possibility of different set of genes playing biological roles to TNB trait. Furthermore, the network analysis across the herd year month (HYM) levels illustrated genes interactions that were consistent with the known mammal's fertility biology and captured known transcription factor (TF), pointing out different sets of candidate genes for TNB for each of the HYM groups. Based on these results, the genomic diversity across environments should be taken into account in breeding programs in such a way that environment specific reference populations should be considered for genomic selection.

Introduction

Reproductive traits such as total number born (TNB) are highly relevant in pig breeding (Bergsma et al., 2008; Knauer et al., 2010) and a large proportion of its

phenotypic variance affected by the environment (Knap and Su, 2008). Nowadays, assisted reproductive techniques, such as artificial insemination, have allowed sire distribution across multiple environments highlighting the relevance of genotype x environment interaction (GxE) in pig production. GxE studies, using genomic information, are important in view of the promising future of genomic wide selection (GWS) aiming to increase the genetic gain in pigs. These analyses can be accessed using different methodologies, such as multitrait models and its generalizations (Meyer, 2009), and also random regression reaction norm models (Kolmodin et al., 2002, Calus and Veerkamp, 2003; Cardoso and Tempelman, 2012).

Under a random regression reaction norm (RRRN) models approach, it is possible to estimate the additive genomic breeding values (GEBV) for animals over different environmental levels, and furthermore it is also possible to estimate single nucleotides polymorphism (SNP) effects over such levels (Silva et al., 2014). These approaches, allow us to perform a genome wide association study (GWAS) defining relevant SNP in each level. Thus, genetic dissection of complex phenotypes, such as TNB, through candidate genes network derived from relevant SNP in each level might provide a better biological understanding of the trait across different environments.

Gene networks from GWAS can be performed using genes related with identified SNP to examine the sharing of pathways and functions involving these genes. In addition, transcription factor (TF) analyses can be implemented. It has been shown that TFs have relations with important traits in pigs, besides their roles to regulate gene expression, e.g. *PIT1* for carcass traits and *SREBF1* related with regulation of muscle fat deposition (Yu et al., 1995 and Chen et al., 2008). Providing evidence for the interaction between TFs and their predicted targets with a further gene-TF network construction must be helpful to point out the most probable group of

candidate genes. There are relevant studies involving gene networks in puberty related traits for cattle (Fortes et al., 2011; Reverter and Fortes, 2013); in pigs this approach has began to be exploited (Verardo et al., 2015).

The objective of this work was to present gene network analysis for total number born trait (TNB) in pigs across different environment using results from a RRRN model GWAS approach. Under this framework, a combined network aiming to obtain a set of candidate genes based on their shared TF binding sites were construct. Therefore, we were able to observe the similarities (or not) at marker and gene level across different environment groups.

Materials and Methods

Ethics Statement

The data used for this study was obtained as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purposes of the breeding program. Data recording and sample collection were conducted strictly in accordance with the Dutch law on the protection of animals (Gezondheids- en Welzijnswet voor Dieren).

Phenotypic and Genotypic Data

Phenotypic data consisted of TNB information from 80794 litters of sows originated from dam lines of the Topigs Norsvin breeding program, daughters of 340 purebred genotyped sires (Large White based). In order to explore the maximum of environments divergences, target sires, with a larger number of daughters spread over a large number of herd-year-month (HYM) levels (n=2764), including different countries were used. The number of parities per sow ranged from 1 to 14. TNB records ranged from 7 to 23.

The sires were genotyped using Illumina PorcineSNP60 Beadchip (San Diego, CA, USA, Ramos et al., 2009). The DNA was prepared from EDTA blood, hair roots or meat samples, using the Gentra Puregene DNA Preparation Kit (Minneapolis, MN), according to the manufacturer's instructions. The extraction was based on a modified salt precipitation method. DNA concentration was measured on the Nanodrop ND-1000 Spectrophotometer (NanoDrop Technologies, LLC, Wilmington, Delaware) and the DNA quality checked by agarose gel.

Positions of the SNPs were based on pig genome build10.2 (Groenen et al., 2012). Cleaning of the SNP data consisted of exclusion of SNPs with: 1) GenCall score <0.15, 2) location on sex chromosomes, 3) call rate <95%, 4) minor allele frequency (MAF) <0.01 and 5) no physical mapping on pig genome build 10.2. After these quality control measures, 44383 SNPs remained for GWAS. All animals showed frequencies of missing genotypes <0.05.

Statistical models

The two-step reaction norm approach (Calus et al., 2002; Kolmodin et al., 2002) was used in order to access the GxE analysis. In agreement with Cardoso and Tempelman (2012), it is the most commonly used approach, in which the first step provides estimates of the environmental effects on phenotype using a general model ignoring G×E, and the second step provides estimates of intercept and slope using a random regression model whose the covariate values are the environmental effects estimates from the first step.

The first step consisted in the fitting of sire model, which was given by:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta}_{\mathbf{cycle}} + \mathbf{X_2}\boldsymbol{\beta}_{\mathbf{hym}} + \mathbf{Zu} + \mathbf{e} \quad [1]$$

where: $\mathbf{y}$ is the vector of TNB for the daughters of the considered sires; $\boldsymbol{\beta}_{\text{cycle}}$ and $\boldsymbol{\beta}_{\text{hym}}$ are, respectively, the fixed effects of cycle and herd-year-month (HYM), whose the

incidence matrices are $\mathbf{X_1}$ and $\mathbf{X_2}$, $\mathbf{u}$ is the vector of sire additive genetic effect, $\mathbf{u}\sim N(0, \sigma_u^2\mathbf{G})$, and $\mathbf{e}$ is the residual random term, $\mathbf{e}\sim N(0, \sigma_e^2\mathbf{I})$.

The aims of this first step were provide a vector of pre-corrected phenotype ($\mathbf{y}^*$) for fixed effects, $y^* = y - \left(\mathbf{X_1\beta_{cycle}} + \mathbf{X_2\beta_{hym}}\right) = \hat{e} + \mathbf{Z}\hat{u}$, and a vector of HYM levels estimates ($\widehat{\boldsymbol{\beta}}_{\mathbf{hym}}$), which are respectively used in the second step as the dependent and independent variables in the reaction norm model.

In the second step, the following random regression reaction norms (RRRN) model was adopted:

$$y_{ij}^* = \mu + a_i + b\hat{\beta}_{hym_{ij}} + \varepsilon_{ij} \qquad [2]$$

where: y^*_{ij} is the pre-corrected TNB values from daughters of sire i in the level j of HYM estimated effect ($\hat{\beta}_{hym}$), μ is the general mean, a_i and b_i are, respectively, the random intercept and random slope for the regression of additive genetic value (u_i) over HYM levels, and ε_{ij} is the residual term, $\varepsilon_{ij}\sim N(0, \sigma_k^2)$. In order to consider the heterogeneity of residual variance, four different classes were assumed: k=1 ($\hat{\beta}_{hym} \leq 13$), 2($13<\hat{\beta}_{hym} \leq 14$), 3($14<\hat{\beta}_{hym} \leq 15$) and 4($\hat{\beta}_{hym}>15$), in which the number of records were, respectively, 11940, 25158, 33029 and 10667. It was assumed also that the joint distribution for $\mathbf{a}$=[a_1, a_2, …, a_{113}]' and $\mathbf{b}$=[b_1, b_2, …, b_{113}]', being $\boldsymbol{\theta} = [\mathbf{a}, \mathbf{b}]'$, was given by: $\theta\sim N(0, \Sigma_{ab} \otimes G)$

Where $\Sigma_{ab} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}$

Under the genomic approach, the **G** matrix was used in the models [1] and [2], being $\mathbf{G = MM'/\Sigma_{m=1}^{M} 2p_m(1 - p_m)}$, where **M** is the incidence SNP matrix assuming 0, 1 and 2, respectively to genotypes aa, aA and AA, and p_m is the allele frequency of SNP m. The models [1] and [2] were fitted by ASREML (Gilmour et al., 2002) software using the *.GIV* qualifier in order to enter with $\mathbf{G^{-1}}$ in the mixed model equations.

For this approach using **G** matrix, the predicted additive genetic effect of sire i in the HYM level j ($\hat{u}_{ij}$) for TNB were given respectively by:

$$\hat{u}_{ij} = \mathbf{K'}_j\widehat{\boldsymbol{\theta}}_i = \hat{a}_i + \hat{b}_1\hat{\beta}_{hym_j}\ , \qquad [3]$$

Using **G** matrix, it was possible to calculate a vector of SNPs effects ($\widehat{\Phi}_j$) for each HYM level. Once the estimate of genomic breeding value (GEBV) of sire i in each level j of HYM is given by [3], being $u_j = \hat{a} + \hat{b}_1\hat{\beta}_{hym_j} = [\hat{u}_{1j}, \hat{u}_{2j}, \ldots, \hat{u}_{340j}]$ the vector of EGBV in the level j of HYM, and assuming, in the simplest way as possible, that $u_j = \mathbf{X\Phi}_j$ (Van Raden, 2008), the normal equation system can be used to provides $\widehat{\Phi}_j$:

$$\widehat{\Phi}_j = (\mathbf{M'M})^-(\mathbf{M'}\hat{u}_j) = (\mathbf{M'M})^-\left(\mathbf{M'}\left(\hat{a} + \hat{b}_1\hat{\beta}_{hym_j}\right)\right) = (\mathbf{M'M})^-\left(\mathbf{M'}\hat{a} + \hat{b}_1\hat{\beta}_{hym_j}\right)$$

$$= \underbrace{(\mathbf{M'M})^-(\mathbf{M'M}\hat{a})}_{\text{SNP effect for a: } \hat{a}_{SNP}} + \underbrace{(\mathbf{M'M})^-(\mathbf{M'}\hat{b})}_{\text{SNP effect for a: } \hat{b}_{SNP}}\ \beta_{hym_j}$$

Thus, a general linear prediction equation can be obtained:

$$\Phi_j = \hat{a}_{SNP} + \hat{b}_{SNP}\hat{\beta}_{hym_j},$$

which allows to estimate a vector of SNP effects for each HYM level of interest, within the observed range of $\hat{\beta}_{hym}$.

The variance explained by each SNP in each HYM level ($\sigma^2_{SNP_j}$) was estimated as follows:

$$\sigma^2_{SNP_j} = 2pq(\Phi_j{}^2)$$

where p and q are the allele frequencies, and the $\widehat{\Phi}_j$ are the SNPs effects for each HYM level. To assess whether SNP have the same effect across HYM levels, we identified the top 0.1% of SNP (44) in each level that showed the greatest variance. The number of shared SNP between the top 0.1% by two HYM levels was used to reflect the similarity of SNP associations across environments.

Linkage disequilibrium (LD) relations across the top 0.1% SNPs were evaluated through Haploview (Barret et al., 2005) to identify QTLs regions (blocks) among chromosomes using the default parameters based on Gabriel et al. (2002) and D prime (Lewontin, 1964).

Network analyses

Genes overlapping single markers plus 25.5 Kbp 5' and 3' flanking sequence, (e.i. the half of the average distance between two SNPs on the Chip) were identified at the reference genome (build10.2). By including the 25.5 Kbp intervals, we verified the presence of any gene that could be linked to a block or a relevant SNP for TNB, since distances between two SNPs around 200-250 Kb, has been demonstrated high LD (r2) values (0.31) in studies of commercial pigs (Veroneze et al., 2014). Aiming to compare the set of genes across HYM levels, we divided the levels in three groups, Low (HYM<12), Mean (12≤HYM<15) and High (HYM≥15) based on the HYM data distribution. The ClueGO Cytoscape plug-in (Bindea et al., 2009) was used to construct gene networks highlighting biological roles and relations across those three set of genes.

The identified sets of genes for each group were submitted to the TFM-Explorer program (http://bioinfo.lifl.fr/TFM/TFME/) aiming to identify the TF related with them. This program takes a set of gene sequences, and searches for locally overrepresented transcription factor binding sites (TFBS) using weight matrices from JASPAR database (Sandelin et al., 2004) to detect all potential TFBS, and extracts significant clusters (region of the input sequences associated with a factor) by calculating a score function. This score threshold is chosen to give a P-value equal or better to 10^{-3} for each position for each sequence such as described in Touzet and Varré (2007). From each set of genes, excluding ncRNA genes, we collected the 3000

bp upstream and 300 bp downstream flanking sequence from the gene start site based on Sscrofa10.2 assembly at NCBI web site. Data from each HYM group was separately used as input at TFM-explorer and the given list of TF was submitted at the Cytoscape (Shannon et al., 2003) using a Biological Networks Gene Ontology tool (BiNGO) plugin (Maere et al., 2005), to determine which GO terms are significantly overrepresented. Based on biological process overrepresented at BiNGO and through literature review, we were able to select the main TFs related with TNB (key TF) in each group, from which we constructed a gene-TF network.

Aiming to identify the most likely candidate genes, we used the NetworkAnalyzer tool within Cytoscape. Based on the number of TFBS and consequently the number of connections in each node, the most connected genes within the TNB gene-TF network were determined in each HYM group. Finally, the gene-TF networks derived for each group were merged, highlighting the genes and TF in common providing an overview across the HYM groups.

Results

Genome-wide association analysis

To investigate the SNPs that were associated with TNB across HYM levels, we identified the most important SNPs for each HYM. The variance explained by each SNP in each HYM level was estimated showing different patterns across HYM levels being the top larger variances present at the highest level (Figure 1a). The top 0.1% of SNP (n=44) was identified for each HYM based on their SNP variance estimates. The number of overlapping SNP between adjacent HYM levels was relatively high (> 29, 65%), whereas there was no SNP that overlapped between the top 0.1% of SNP at the lowest and highest HYM levels (Figure 2). A total of 127 SNPs were identified spread across all chromosomes (Table S1). From these, six SNPs formed two blocks

on chromosome 10 and four SNPs formed one block at chromosome 3. The variances of the SNPs that formed the Blocks were plotted showing their behavior across levels (Figure 1b, 1c and 1d). From each relevant SNP and Block, we found a total of 79 genes (Table S1).

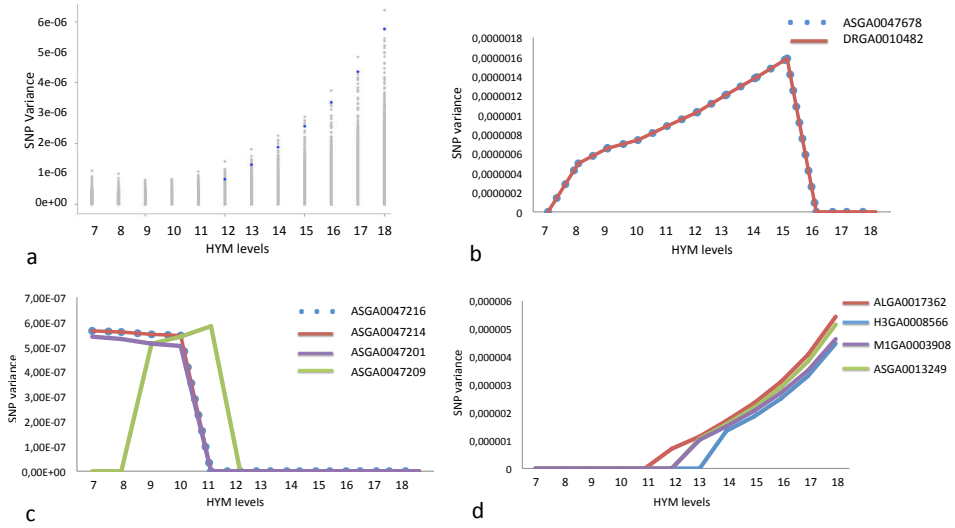


Figure 1 SNPs variance across HYM levels
(a) The variance explained by each SNP across all 12 HYM levels highlighting a specific SNP behavior (Blue dot) that have the highest variance and was related with a gene (C1QTNF9). (b) The Block 1 behavior across the levels composed by ASGA0047678 and DRGA0010482 SNPs variance. (c) The Block 2 behavior across the levels composed by ASGA0047216, ASGA0047214, ASGA0047201 and ASGA0047209 SNPs variance, on this case, the block was present only at HYM 9 and 10. (d) The Block 3 behavior across the levels composed by ALGA0017362, H3GA0008566, M1GA0003908 and ASGA0013249 SNPs variance.

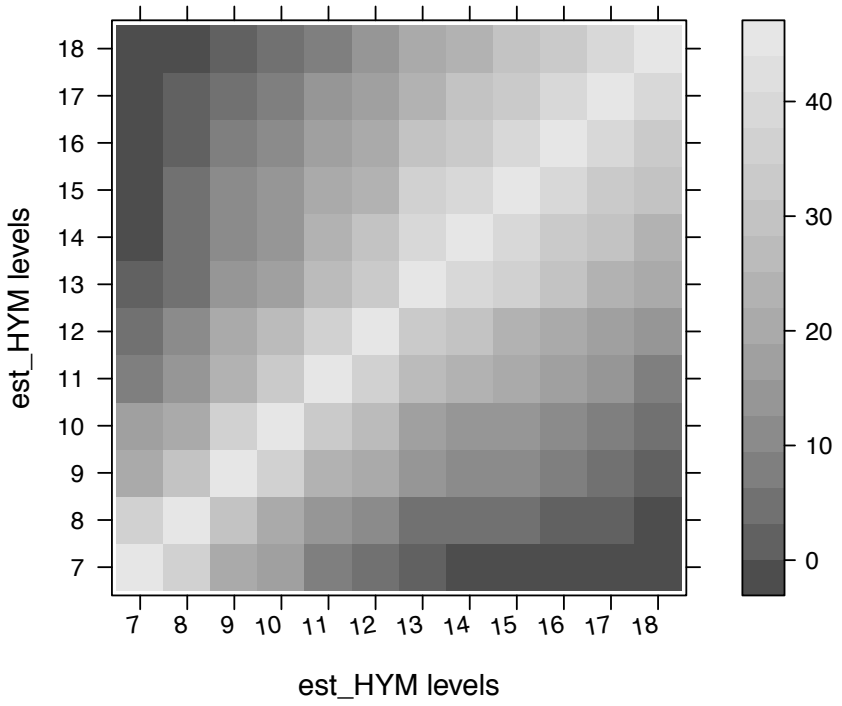


Figure 2 Top 0,1% overlapping genes
Number of common SNPs between the 44 most relevant SNPs at each HYM level.

Network analyses

The ClueGO Cytoscape plug-in was used to construct gene networks highlighting biological roles and relations across the three set of genes, Low (HYM<12), Mean (12≤HYM<15) and High (HYM≥15). On this step, an important biological process for TNB (ovulation cycle process) was identified being shared by four genes: BMPR1B present at the Low group; TGFB3 and SLIT3 present at the High group and NCOR2 present in the Mean and High groups (Figure 3). In addition, regulatory sequence analyses were performed for all detected genes.

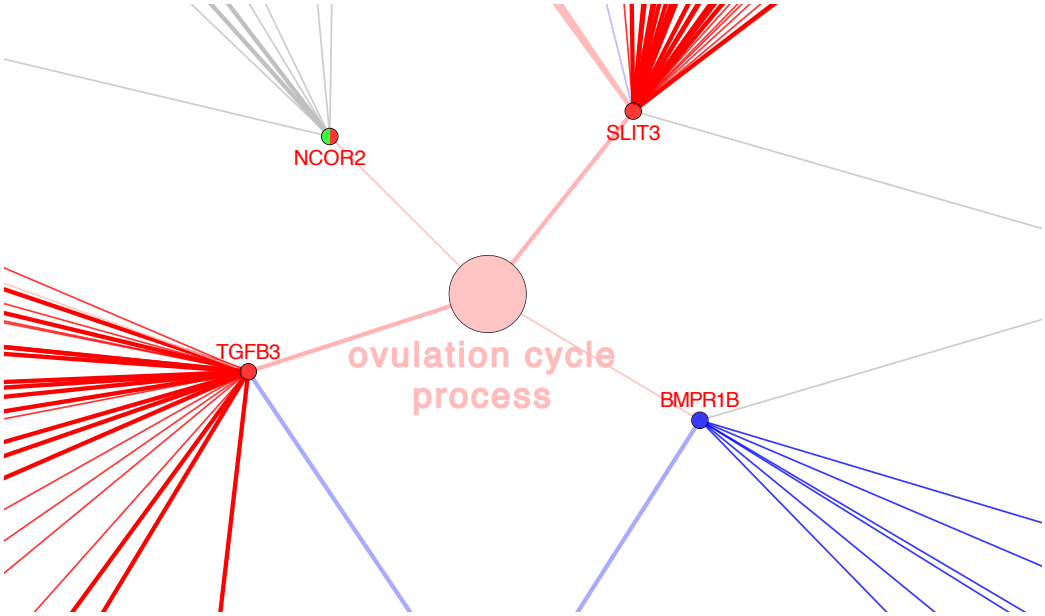


Figure 3 Main biological process networks
Functionally group network with a main biological process term (pink node) and genes (labeled in red). Blue nodes represent genes from Low group, red nodes from High group and green and red node is a gene present at the Mean and High group. The node size represents the term enrichment significance from ClueGO.

Each group of genes was separately used as input at TFM-explorer. A total of 19, 19 and 20 TFs were identified for the Low, Mean and High groups respectively. Based on biological process overrepresented at BiNGO and through literature review, we were able to select the main TFs related with TNB (key TF) in each group (Table 1), from which we constructed a gene-TF network highlighting the most connected genes (Figure 4).

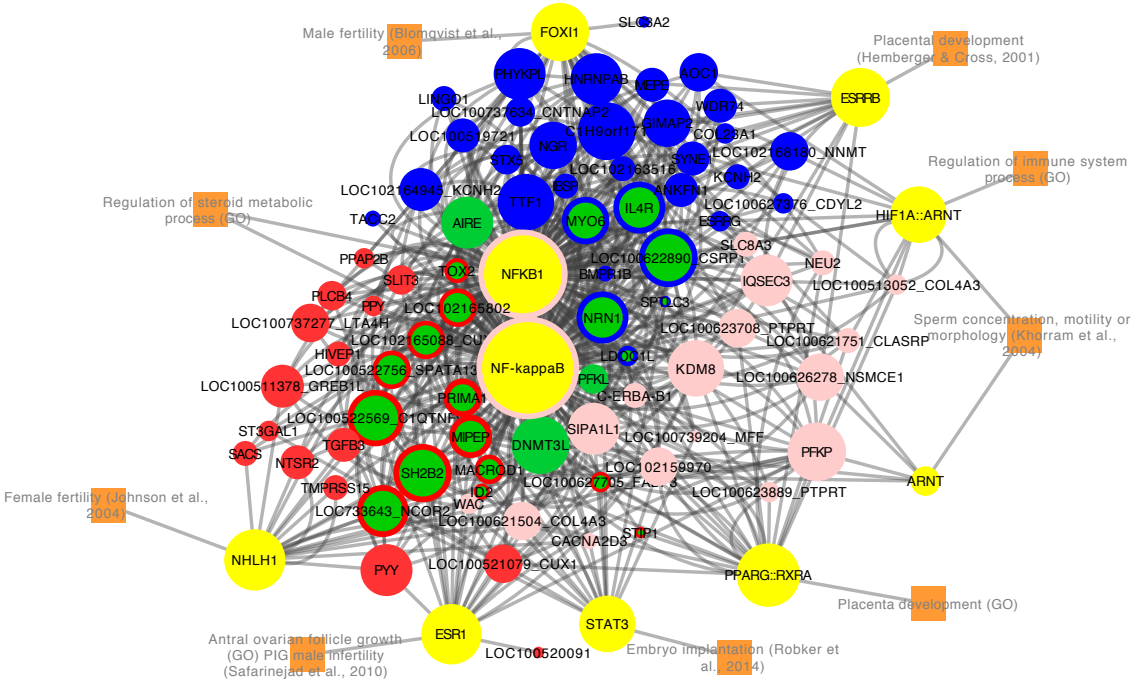


Figure 4 Gene-Transcription Factor (TF) network

Genes related with relevant SNPs for TNB in each group (blue nodes, Low group; green nodes, Mean groups; red nodes, High group) and in common to all groups (pink nodes). Green nodes with blue or red border are genes in common to Low or High with the Mean group respectively. Associated with these genes, we have TFs (yellow nodes), yellow node with pink border are in common to all groups). The node size corresponds to the network analyses (Cytoscape) score where bigger nodes represent higher edges

Table 1 – Main Transcription Factors
Most representative transcription factors (TF), associated with genes identified for the groups Low, Mean and High, based on their biological process and literature evidences.

TF	Groups	Biological Process	Literature evidences
NFKB1	Low, Mean and High	Regulation of steroid metabolic process	Endometriosis and idiopathic infertility (Bianco et al., 2012)
NF-kappaB	Low, Mean and High	Regulation of steroid metabolic process	Regulation of endometrial receptivity (Song et al., 2012)
HIF1A::ARNT	Low	Regulation of immune system	Placental development (Letta et al., 2006) and sperm function (Khorram et al., 2004)
ESRRB	Low	Regulation of biological process	Placental development (Hemberger & Cross 2001)
FOXI1	Low and High	Embryonic development	Male fertility (Blomqvist et al., 2006)
PPARG::RXRA	Mean	Placenta development	Sperm metabolism in pigs (Santoro et al., 2013) and placentogenesis (Wendling et al., 1999)
STAT3	Mean and High	Response to estradiol stimulus	Embryo implantation (Robker et al., 2014)
ESR1	Mean and High	Uterus development	Male fertility (Safarinejad et al., 2010)
NHLH1	High	Developmental process	Female fertility (Johnson et al., 2004)
ARNT	High	Embryonic development ending in birth or egg hatching	Sperm function (Khorram et al., 2004)

Discussion

On this work we performed post-GWAS (gene networks) analysis using results from a two-step reaction norm approach in order to access the GxE interaction. We used this approach to estimate different SNP effects for each HYM level. Considering the top 0.1% of SNP for each HYM based on their SNP variance estimates, a total of 127 SNPs were identified spread across all chromosomes. According to Silva et al. (2014), we observed that the variance of the top SNPs varied across HYM levels showing higher values at high HYM levels compared to low HYM levels. Following the SNP variance, the three Blocks identified on the present study were present during specific levels. The first Block (Figure 1b) was present over all HYM groups (Low, Mean and High) while Blocks two and three (Figure 1c and 1d) were present mainly at the Low and High group respectively. The highlighted structure illustrates the changing of

genome regions relevance for the trait across distinct environments. In addition, we could notice that the number of overlapping SNP between two adjacent HYM levels was relatively high (> 65%), whereas there was no SNP that overlapped between the top 0.1% of SNP at the lowest and highest HYM levels. Based on that, it seems that different genes play distinct biological roles related to the trait disclosing different sets of relevant SNPs across each level. Considering these outcomes, we search for related genes to perform post-GWAS analyses.

Considering HYM groups and based on each relevant SNP and Block, we found a total of 79 genes. A gene network highlighting the biological roles and relations across the sets of genes from each group was constructed. On this step, an important biological process for TNB (ovulation cycle process) was identified being shared by four genes: BMPR1B present at the Low group; TGFB3 and SLIT3 present at the High group and NCOR2 present in Mean and High groups. BMPR1B is a bone morphogenetic receptor type 1 B gene and has been related with the Booroola phenotype in sheep (Souza et al., 2001), characterized by a high prolificacy. Interestingly, on this work this gene is found at the Low group suggesting an opposite way of this gene to act in pigs. This is reinforced by Tomás et al. (2006) in an association study between polymorphisms of BMPR1B and reproductive performance of Iberian X Meishan F2 sows that found negative additive genetic effect estimates of this gene for TNB.

Also related with ovulation cycle process at the network, we identified the TGFB3, SLIT3 and NCOR2. TGFB3 is an isoform of transforming growth factor beta. It has been suggested that TGFB3 may play an important role in modulating theca cell function of pre- and postovulatory follicles of the pig (Steffl et al., 2008). At the same group of TGFB3, SLIT3 also plays a role in reproductive systems. Studies suggest

that this gene, among others from the same family, could play an important role in luteolysis in women and also may influence processes that occur during early ovary development such as follicle formation or oocyte survival (Dickinson and Duncan, 2010, Dickinson et al., 2008 and Dickinson et al., 2010). The fourth gene, NCOR2 is a nuclear receptor co-repressor 2 cited to be related with ovarian follicular cysts (Salvetti et al., 2012). In sows, the presence of ovarian cysts has been associated with a greater return to estrus rate, a decreased farrowing rate, and an increased percentage of anestrus (Castagna et al., 2004).

The biological process network in which these four genes (BMPR1B, TGFB3, and NCOR2) were inserted was our first evidence that genes, related to SNPs found at low levels, are playing a role on TNB as well as those estimated at high HYM levels. Besides, we performed a gene-TF network aiming to identify additional genes associated with TNB in different levels. To build the gene-TF network we first selected key TF connected with the three sets of genes (Low, Mean and High). These TF were related with important biological process (e.g. regulation of steroid metabolic process, embryonic and placenta development) and literature evidences on male and female fertility as observed in Table 1. Based on these key TFs we were able to build a gene-TF network highlighting genes sharing roles with TNB across all groups (e.g. HNRNPAB, DNMT3L, PYY and PFKP).

The HNRNPAB gene is found at Low group being one of the most highlighted considering the large number of connections to key TFs. This gene has been indicated to play an important role in spermatogenesis by regulating stage-specific translation of testicular mRNAs (Fukuda et al., 2013). Belonging to the Mean group, DNMT3L is a member of the DNA methyltransferase 3 family and is also found to be involved with spermatogenesis in mouse (Webster et al., 2005). Methylation has been cited to

occur during oogenesis and spermatogenesis (Sanford et al., 1987) and may have an important role in early lineage specification on mouse embryos (Santos et al., 2002).

The third gene, PYY, belongs to the High group being cited by Fernandez-Fernandez et al. (2006) to play a role in the neuroendocrine networks integrating energy balance and reproduction. The PFKP gene is present in all groups. This gene has been found to be involved in glycolysis and highly expressed in cumulus cells of mouse antral follicles (Sugiura et al. 2005). Also, proposed to identify oocytes with high developmental potential, leading to enhanced implantation rates and greater developmental capacity throughout gestation (Gebhardt et al., 2011). Furthermore, in all groups we identified others well connected genes with key TFs.

At Low group, others highlighted genes were found (e.g. PHYKPL, GMAP2, TTF1 and BMPR1B). These genes were linked with TFs related with placenta development and male fertility as ESRRB and FOXI1 respectively (Hemberger & Cross 2001 and Blomqvist et al., 2006). Belonging to the Mean group, we can see also other genes as AIRE, IL4R, CSRP1, NRN1, SH2B2 and NCOR2. These genes were well connected with NFKB1, NF-KappaB and PPARG::RXRA TFs which are related with endometriosis, regulation of endometrial receptivity, sperm metabolism in pigs and placentogenesis in mouse (Bianco et al., 2012, Song et al., 2012, Santoro et al., 2013 and Wendling et al., 1999). At High group, we identified genes as C1QTNF9, GREB1L, LTA4H, TGFB3 and SLIT3. In our study, these genes are linked with NHLH1 and ESR1 TFs related with female and male fertility respectively (Johnson et al., 2004 and Safarinejad et al., 2010). In addition, the C1QTNF9 gene, as well as the PYY present on this group, is also related with energy balance (Wei et al., 2014). This gene is marked by the second top SNP from the highest level (see Figure 1a) of the High group. The evidence of energy related genes showing SNPs with a large

variance indicates the importance of energy balance role for TNB trait mainly on high production group.

Interestingly, as noted for each group, in general a diverse set of genes was identified but overlapping biological roles regarding male or female fertility. Thus, for each HYM group selection for the same trait has led to recruitment of different genes involved in distinct biological processes delivering the TNB phenotypes. The explanation for selection on different set of genes is still not fully understood but recent publications have found similar observation. Machado et al. (2010) in a QTL study for tick resistance/susceptibility in a bovine F2 population found that depending on the tick evaluation season (dry and rainy), different sets of genes could be involved in the tick resistance mechanism. Working with commercial pigs, Veroneze et al. (2014), observed differences in linkage disequilibrium patterns and persistence of phase across pig lines suggesting different QTL and distinct set of genes in each pig line.

Furthermore, in a GxE interaction study using the same approach applied in the present research, Silva et al. (2014) observed also variation in the significant SNP set for total number of piglets born in a commercial line according to environment variations resulting in different QTLs. The gene-TF networks analysis in this study, further illustrate the phenomenon observed in these previous studies, showing that the same trait can be under selection without selecting for the same set of genes. Selection can act on many genes and depending on the environment (variants present in that) initial selection can be on different genes and this may directly have an influence on what genes are under selection at later stages. Alternatively, different genes involved in distinct biological pathways may illustrate the physiological requirement according

to environment level (e.g. HNRNPAB at the Low group and PYY and C1QTNF9 genes present at the High group).

Thus, different components of pig physiology (e.g. reproductive and endocrine systems) may be important for determining the final number of piglets born, and different genes are therefore important and contribute to the observed genetic differences. Analyses of data from different environment are valuable for animal breeding as they provide additional insights into the genes controlling the trait of interest under distinct circumstances. From our study, we conclude that within the breeding programs on these different environments, specific sets of genes have been selected explaining the observed different set of genes involved in the same trait.

Conclusion

The post-GWAS performed on this study provided by gene network analyses generated a rich information resource about genes identified in a random regression reaction norm models GWAS approach for TNB across different environments (HYM levels). Different sets of relevant SNPs and QTL blocks across the HYM levels were identified leading to the possibility of different set of genes playing biological roles to TNB trait (e.g. spermatogenesis, placentogenesis and energy balance). Furthermore, the network analysis across the HYM levels illustrated genes interactions that were consistent with the known mammal's fertility biology and captured known TF, pointing out different sets of candidate genes for TNB in each of the HYM groups. Based on these results, the genomic diversity across environments should be taken into account in breeding programs in such a way that environment specific reference populations should be considered for genomic selection. Moreover, a large number of genes and transcription factors interactions leading to complex traits as total number born still to be identified highlighting the importance on post-GWAS analyses.

Reference

Bergsma R, Kanis E, Verstegen MWA, Knol EF: **Genetic parameters and predicted selection results for maternal traits related to lactation efficiency in sows**. *J Anim Sci*, 2008, **86**:1067-1080

Knauer MT, Cassady JP, Newcom DW, See MT: **Phenotypic and genetic correlations between gilt estrus, puberty, growth, composition, and structural conformation traits with first litter reproductive measures**. *J Anim Sci* 2010, **89**:935:942.

Knap PW, Su G: **Genotype by environment interaction for litter size in pigs as quantified by reaction norms analysis**. *Animal* 2008, **2**:1742–1747.

Meyer K: **Factor-analytic models for genotype × environment type problems and structured covariance matrices**. *Genet Sel Evol* 2009, **41**:21–32.

Kolmodin R, Strandberg E, Madsen P, Jorjani H: **Genotype by environment interaction in Nordic dairy cattle studied using reaction norms**. *Acta Agric Scand* 2002, **52**:11–24.

Calus MPL, Veerkamp R: **Estimation of environmen- tal sensitivity of genetic merit for milk production traits using a random regression model**. *J Dairy Sci* 2003, **86**:3756–3764.

Cardoso FF, Tempelman RJ: **Linear reaction norm models for genetic merit prediction of Angus cattle under geno- type by environment interaction**. *J Anim Sci* 2012, **90**(7):2130–2141.

Silva FF, Mulder HA, Knol EF, Lopes MS, Guimarães SEF, Lopes PS et al.: **Sire evaluation for total number born in pigs using a genomic reaction norms approach**. *J Anim Sci* 2014, **92**(9), 3825-3834.

Yu TP, Tuggle CK, Schmitz CB, Rothschild MF: **Association of PIT1 polymorphisms with growth and carcass traits in pigs**. *J Anim Sci* 1995, **73**(5):1282-1288.

Chen J, Yang XJ, Xia D, Wegner J, Jiang Z, Zhao RQ: **Sterol regulatory element binding transcription factor 1 expression and genetic polymorphism significantly affect intramuscular fat deposition in the longissimus muscle of Erhualian and Sutai pigs**. *J Anim Scie* 2008, **86**(1):57-63.

Fortes MR, Reverter-Gomez T, Hiriyur-Nagaraj S, Zhang Y, Jonsson N, Barris W, et al.:**A SNP-derived regulatory gene network underlying puberty in two tropical breeds of beef cattle**. *J Anim Sci* 2011, **89**:1669-1683.

Reverter A, Fortes MRS: **Building single nucleotide polymorphism-derived gene regulatory networks: Towards functional genomewide association studies**. *J Anim Sci* 2013, **91**:530-536.

Verardo LL, Silva FF, Varona L, Resende MDV, Bastiaansen JWM, Lopes PS, et al.: **Bayesian GWAS and network analysis revealed new candidate genes for number**

of teats in pigs. J App Genet 2014, 1-10.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE et al.: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology**. *PLoS ONE* 2009, **4**:e6524.

Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF et al.: **Analyses of pig genomes provide insight into porcine demography and evolution**. *Nature* 2012, **491**(7424):393-398.

Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R: **ASReml user guide release 1.0**. VSN Int. Ltd. 2002, Hemel Hempstead, UK.

Van Raden P: **Efficient methods to compute genomic predictions**. J Dairy Sci 2008, **91**:4414–4423.

Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**:263-265.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al.: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**:2225-9.

Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models**. *Genetics* 1964, **49**:49-67.

Bindea G, Mlecnik B, Hack H, Charoentong P, Tosolini M, Kirilovsky A, et al.: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics* 2009, **25(**8):1091-1093.

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, (32 Database):D91-4.

Touzet H, Varré JS: **Efficient and accurate P-value computation for Position Weight Matrices**. *Algorithms Mol Biol* 2007, **2**(1510.1186):1748-7188.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al.: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genom Res* 2003, **13**(11):2498-2504.

Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448-3449.
Souza, CJ, MacDougall C, Campbell BK, McNeilly AS, Baird DT: **The Booroola (FecB) phenotype is associated with a mutation in the bone morphogenetic receptor type 1 B (BMPR1B) gene**. *J Endocrinol* 2001, **169(**2):R1-R6.

Tomas A, Frigo E, Casellas J, Ramirez O, Ovilo C, Noguera JL et al.: **An association study between polymorphisms of the porcine bone morphogenetic protein receptor type1β (BMPR1B) and reproductive performance of Iberian× Meishan F2 sows**. *Anim Genet* 2006, **37**(3):297-298.

Steffl M, Schweiger M, Amselgruber WM: **Expression of transforming growth factor-beta3 (TGF- beta3) in the porcine ovary during the oestrus cycle**. *Histol Histopathol* 2008, 23:665–671.

Dickinson RE, Duncan WC: **The SLIT–ROBO pathway: a regulator of cell function with implications for the reproductive system**. *Reproduction* 2010, **139**(4):697-704.

Dickinson RE, Myers M, Duncan WC: **Novel regulated expression of the SLIT/ROBO pathway in the ovary: possible role during luteolysis in women**. *Endocrinol* 2008, **149**:5024–5034.

Dickinson RE, Hryhorskyj L, Tremewan H, Hogg K, Thomson AA, McNeilly AS et al.: **Involvement of the SLIT/ROBO pathway in follicle development in the fetal ovary**. *Reproduction* 2010, **139**:395–407.

Salvetti NR, Alfaro NS, Velázquez MM, Amweg AN, Matiller V, Díaz, PU et al.: **Alteration in localization of steroid hormone receptors and coregulatory proteins in follicles from cows with induced ovarian follicular cysts**. *Reproduction* 2012, **144**(6):723-735.

Castagna CD, Peixoto CH, Bortolozzo FP, Wentz I, Neto GB, Ruschel F: **Ovarian cysts and their consequences on the reproductive performance of swine herds**. *Anim Reprod Sci* 2004, **81**:115-123.

Fukuda N, Fukuda T, Sinnamon J, Hernandez-Hernandez A, Izadi M, Raju CS et al.: **The transacting factor CBF-A/Hnrnpab binds to the A2RE/RTS element of protamine 2 mRNA and contributes to its translational regulation during mouse spermatogenesis**. *PLoS Genet* 2013, **9**(10):e1003858.

Hemberger M, Cross JC: **Genes governing placental development**. *Trends Endocrinol Metab* 2001, **12**(4):162-168.

Blomqvist SR, Vidarsson H, Söder O, Enerbäck S: **Epididymal expression of the forkhead transcription factor Foxi1 is required for male fertility**. *The EMBO journal* 2006, **25**(17):4131-4141.

Sanford JP, Clark HJ, Chapman VM, Rossant J: **Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse**. *Genes & development* 1987, **1**(10):1039-1046.

Santos F, Hendrich B, Reik W, Dean W: **Dynamic reprogramming of DNA methylation in the early mouse embryo**. *Developmental biology* 2002, **241**(1):172-182.

Fernandez-Fernandez R, Martini AC, Navarro VM, Castellano JM, Dieguez C, Aguilar E et al.: **Novel signals for the integration of energy balance and reproduction**. *Mol Cell Endocrinol* 2006, **254**:127-132.

Wei Z, Lei X, Petersen PS, Aja S, Wong GW: **Targeted deletion of C1q/TNF-related protein 9 increases food intake, decreases insulin sensitivity, and promotes hepatic steatosis in mice**. *Am J Physiol Endocrinol Metab* 2014, **306**(7):E779-E790.

Johnson SA, Marín-Bivens CL, Miele M, Coyle CA, Fissore R, Good DJ: **The Nhlh2 transcription factor is required for female sexual behavior and reproductive longevity**. *Horm Behav* 2004, **46**(4):420-427.

Safarinejad MR, Shafiei N, Safarinejad S: **Association of polymorphisms in the estrogen receptors alpha, and beta (ESR1, ESR2) with the occurrence of male infertility and semen parameters**. *J Steroid Biochem Mol Biol* (2010), **122**(4):193-203.

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J et al.: **Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis**. *Proc Natl Acad Sci USA* 2005, **102**(11):4068-4073.

Bianco B, Lerner TG, Trevisan CM, Cavalcanti V, Christofolini DM, Barbosa CP: **The nuclear factor-kB functional promoter polymorphism is associated with endometriosis and infertility**. *Hum Immunol* 2012, **73**(11):1190-1193.

Song Y, Wang Q, Huang W, Xiao L, Shen L, Xu W: **NF-kappaB expression increases and CFTR and MUC1 expression decreases in the endometrium of infertile patients with hydrosalpinx: a comparative study**. *Reprod Biol Endocrinol* 2012, **10**:86.

Santoro M, Guido C, De Amicis F, Sisci D, Vizza D, Gervasi S et al.: **Sperm metabolism in pigs: a role for peroxisome proliferator-activated receptor gamma (PPARγ)**. *J Exp Biol* 2013, **216**(6):1085-1092.

Wendling O, Chambon P, Mark M: **Retinoid X receptors are essential for early mouse development and placentogenesis**. *Proc Natl Acad Sci USA* 1999, **96**(2):547-551.

Sugiura K, Pendola FL, Eppig JJ: **Oocyte control of metabolic cooperativity between oocytes and companion granulosa cells: energy metabolism**. *Dev Biol* 2005, **279**(1):20-30.

Gebhardt KM, Feil DK, Dunning KR, Lane M, Russell DL: **Human cumulus cell gene expression as a biomarker of pregnancy outcome after single embryo transfer**. *Fertil Steril* 2011, **96**(1):47-52.

Machado MA, Azevedo AL, Teodoro RL, Pires MA, Peixoto MG, de Freitas C, et al.: **Genome wide scan for quantitative trait loci affecting tick resistance in cattle (Bos taurus x Bos indicus)**. *BMC Genomics* 2010, **11**:280.

Veroneze R, Bastiaansen J, Knol EF, Guimarães S, Silva FF, Harlizius B, et al.: **Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations**. *BMC Genet* 2014, **15**(1):126.

Additional file

Additional file 1 – Table S1. Significant SNPs for total number born and associated genes. The table shows the relevant SNPs (top 0,1%), their positions in base pairs (bp) at swine chromosome (Chr), their belonged group (Low, Mean and High), the QTL Block, associated genes (genes located in the Block or in an interval of 25,5 Kbp around each Block or SNP) followed by their distance in base pair of a single marker or in relation to the first or last SNP of the block

SNP	chr	pb	HYM	Block	Genes	Distance (bp)
H3GA0000774	1	13005744	17 and 18	-	-	-
ALGA0000960	1	13253219	14, 15, 16, 17 and 18	-	-	-
ASGA0001200	1	16463222	8 and 9	-	SYNE1	in
ASGA0082812	1	101065951	7, 8, 9, 10, 11 and 12	-	MYO6	in
MARC0088490	1	116134872	16 and 17	-	-	
MARC0008048	1	306269050	7 and 8	-	TTF1	24750
				-	C1H9orf171	in
H3GA0005795	2	6829568	13, 14, 15, 16, 17 and 18	-	-	-
ASGA0008834	2	7058677	12, 13, 14, 15, 16, 17 and 18	-	STIP1	18310
				-	MACROD1	in
				-	LOC102165802	3866
ASGA0093436	2	8270193	7 and 8	-	WDR74	6463
				-	STX5	12820
				-	SLC3A2	in
ASGA0102363	2	81635308	7	-	COL23A1	in
DIAS0001491	2	81800958	7, 8 and 9	-	PHYKPL	1418
				-	HNRNPAB	in
ALGA0114773	2	152813354	16, 17 and 18	-	-	-
H3GA0008566	3	9021679	14, 15, 16, 17 and 18	Block 1*	LOC102165088 /CUX1, CUX1 and SH2B2	in
ALGA0017362	3	9398736	12, 13, 14, 15, 16, 17 and 18			
ASGA0013249	3	9410336	13, 14, 15, 16, 17 and 18			
M1GA0003908	3	9481293	13, 14, 15, 16, 17 and 18		LOC100521079 /CUX1	24380
H3GA0008890	3	15359797	13, 14, 15, 16, 17 and 18	-	-	-
ALGA0017913	3	19801787	9, 10, 11, 12, 13, 14, 15 and 16	-	LOC102159970	3419
ALGA0017926	3	19875986	10, 11 and 12	-	KDM8	in
				-	IL4R	6660
H3GA0008998	3	19921287	9, 10, 11, 12, 13, 14, 15 and 16	-	LOC100626278 /NSMCE1	in
				-	KDM8	in
ASGA0095600	3	20071065	12, 13 and 14	-	-	-

ALGA0107444	3	133863736	16, 17 and 18	- -	LOC100511378 NTSR2	in 4361
H3GA0052445	3	134408795	16, 17 and 18	-	-	-
ASGA0105919	3	135558142	18	-	-	-
ASGA0105169	3	135817713	13, 14, 15, 16, 17 and 18	-	ID2	13450
M1GA0005574	4	7775696	15, 16, 17 and 18	-	ST3GAL1	in
H3GA0015158	5	2021416	7, 8, 9, 10, 11, 12 and 13	-	LDOC1L	15660
ASGA0023762	5	2033140	9, 10, 11 and 12	-	-	-
M1GA0007944	5	69759629	11, 12, 13, 14, 15, 16, 17 and 18	-	IQSEC3	615
ASGA0089191	5	90968236	7, 8, 9 and10	-	-	-
ASGA0083093	5	91636108	18	-	LOC100737277 /LTA4H	8957
ALGA0034199	5	107165464	12, 13 and 14	-	-	-
ASGA0094024	6	7600738	7 and 8	-	LOC100627376 /CDYL2	16230
ASGA0027522	6	7772810	7, 8, 9 and 10	-	-	-
DRGA0006951	6	134411824	7, 8, 9, 10 and 11	-	LOC100519721	in
ASGA0029778	6	139221327	7, 8 and 9	-	-	-
ALGA0116974	6	143740794	15, 16, 17 and 18	-	PPAP2B	6714
ALGA0037499	6	144026033	7, 8 and 9	-	-	-
ASGA0091755	6	157319826	11, 12, 13, 14, 15, 16, 17 and 18	-	-	-
ALGA0037960	7	2255217	7 and 8	-	-	-
ALGA0037893	7	2676514	7 and 8	-	-	-
ALGA0037892	7	2730086	7 and 8	-	-	-
H3GA0019406	7	2909666	7 and 8	-	LOC102163516 /CDYL	in
H3GA0019527	7	3813663	9, 10, 11 and 12	-	NRN1	6413
ALGA0038434	7	8871000	17 and 18	-	HIVEP1	in
M1GA0009528	7	9462748	7	-	-	-
DRGA0007119	7	9493010	7 and 8	-	-	-
ASGA0034251	7	62438211	7	-	LINGO1	in
ASGA0035326	7	98700003	10, 11, 12, 13, 14, 15, 16 and 17	-	-	-
ALGA0043701	7	98760402	9, 10, 11, 12, 13, 14, 15, 16, 17 and 18	-	-	-
H3GA0022580	7	99992288	10, 11, 12, 13, 14, 15, 16, 17 and 18	-	SLC8A3	in
ALGA0043874	7	101215291	11, 12, 13, 14 and 15	-	SIPA1L1	in
ALGA0043984	7	105161320	16, 17 and 18	-	TGFB3 LOC100739559 /TTLL5	17320 in
ASGA0036466	7	121774453	14, 15, 16, 17 and 18	-	PRIMA1	13570

ALGA0045473	7	126374056	7 and 8	-	-	-
ALGA0047109	8	29026480	13, 14 and15	-	-	-
H3GA0024578	8	29624896	9	-	-	-
ASGA0039480	8	109779460	7 and 8	-	-	-
ALGA0049066	8	114008201	7	-	-	-
ALGA0049529	8	134044656	11	-	BMPR1B	in
ALGA0049531	8	134098853	11	-	BMPR1B	in
ALGA0049862	8	140437302	7	-	MEPE IBSP	6793 1137
DRGA0008904	8	140460312	7	-	IBSP	8115
H3GA0027079	9	46701876	7 and 8	-	LOC102168180 /NNMT	3148
H3GA0041349	9	121409854	7, 8, 9, 10 and 11	-	LOC100737634 /CNTNAP2	in
ALGA0056208	10	1725859	7, 8 and 9	-	-	-
H3GA0028897	10	1994574	17 and 18	-	-	-
ALGA0056709	10	8807402	7, 8 and 9	-	ESRRG	in
ASGA0046159	10	8965689	7, 8 and 9	-	-	
H3GA0029253	10	13119118	18	-	LOC100520091	in
H3GA0029481	10	20485599	12, 13, 14, 15, 16, 17 and 18	-	-	-
MARC0034242	10	28408776	8, 9, 10, 11 and 12	-	LOC100622890 /CSRP1	10720
ASGA0047216	10	30450537	7, 8, 9 and 10	Block 2**	LOC100515232	in
ASGA0047214	10	30481283	7, 8, 9 and 10		LOC100515232	24000
ASGA0047209	10	30502415	9, 10 and 11		-	-
ASGA0047201	10	30592523	7, 8, 9 and 10		-	-
ASGA0098394	10	32310120	9, 10, 11, 12, 13 and 14	-	-	-
ASGA0047678	10	44144334	8, 9, 10, 11, 12, 13, 14 and 15	Block 3	LOC100621751 /CLASRP	23240
DRGA0010482	10	44156925	8, 9, 10, 11, 12, 13, 14 and 15		WAC	in
ASGA0047683	10	44272596	8, 9, 10, 11 and 12	-	-	-
MARC0093414	10	44342945	7, 8, 9, 10, 11 and 12	-	-	-
ASGA0047702	10	44617028	7, 8, 9, 10, 11, 12 and 13	-	-	-
ALGA0114416	10	65228340	7 and 8	-	-	-
H3GA0030748	10	72363312	9	-	-	-
ASGA0048983	10	72513357	7, 8, 9, 10, 11 and 12	-	-	-
M1GA0014460	10	73596905	9, 10, 11, 12, 13, 14, 15, 16, 17 and 18	-	PFKP	in
M1GA0014582	11	1762098	15, 16, 17 and 18	-	SACS	17660
H3GA0030943	11	2069845	13, 14, 15, 16, 17 and 18	-	MIPEP	in

ALGA0060277	11	2151176	12, 13, 14, 15, 16, 17 and 18	-	LOC100522756 /SPATA13	8853
					LOC100522569 /C1QTNF9	in
					MIPEP	7102
ALGA0065514	12	19469685	17 and 18	-	PYY	24740
					PPY	14550
ASGA0088869	12	25681693	7 and 8	-	NGR	20470
ASGA0054011	12	32490751	9, 10 and 11	-	-	-
ALGA0065989	12	33348283	9 and 10	-	ANKFN1	in
H3GA0035433	13	12722359	8, 9, 10, 11, 12, 13, 14, 15, 16 and 17	-	C-ERBA-B1	in
ALGA0069669	13	40331249	11, 12, 13, 14, 15, 16, 17 and 18	-	CACNA2D3	5817
H3GA0037747	13	193044253	18	-	TMPRSS15	in
H3GA0054163	13	217027218	12 and 13	-	PFKL	20060
					DNMT3L	866
					AIRE	1002
MARC0025520	14	30477142	13, 14, 15, 16, 17 and 18	-	LOC733643/NC OR2	in
H3GA0042844	14	143051194	7, 8, 9 and 10	-	TACC2	in
ASGA0067289	14	143102976	7, 8, 9 and 10	-	TACC2	in
ALGA0083464	15	4352064	7	-	-	-
ASGA0071140	15	142465360	9, 10, 11, 12, 13, 14, 15 and 16	-	LOC100621504 /COL4A3	in
ASGA0071144	15	142485558	8, 9, 10, 11, 12, 13, 14 and 15	-	LOC100621504 /COL4A3	in
ALGA0087783	15	142546880	9, 10, 11, 12, 13, 14, 15, 16 and 17	-	LOC100513052 /COL4A3	in
					LOC100739204 /MFF	15030
					LOC100621504 /COL4A3	9274
ASGA0083076	15	147552216	10, 11, 12, 13, 14, 15, 16, 17 and 18	-	NEU2	6809
INRA0061113	15	149799164	11, 12, 13 and 14	-	-	-
ALGA0088352	15	151208556	15, 16, 17 and 18	-	-	-
ASGA0071843	16	1016806	7 and 8	-	-	-
H3GA0045756	16	1032442	7, 8, 9 and 10	-	-	-
ALGA0090922	16	59756435	18	-	SLIT3	in
INRA0052808	17	17548564	13, 14, 15, 16, 17 and 18	-	-	-
MARC0013253	17	17686438	14, 15, 16, 17 and 18	-	LOC100627705 /FABP3	2838
H3GA0047972	17	18959392	11, 12, 13, 14, 15, 16, 17 and 18	-	-	
ASGA0075615	17	20488772	15, 16, 17 and 18	-	PLCB4	in

ALGA0093755	17	24374766	10, 11, 12 and 13	-	SPTLC3	in
MARC0037879	17	33561064	10 and 11	-	-	-
MARC0085770	17	33643981	9, 10, 11, 12, 13, 14, 15, 16 and 17	-	-	-
ASGA0094610	17	48307754	11, 12, 13 and 14	-	-	-
MARC0009678	17	49842888	11, 12, 13, 14, 15, 16, 17 and 18	-	LOC100623889 /PTPRT LOC100623708 /PTPRT	24430 in
H3GA0049112	17	50071035	17 and 18	-	LOC100623889 /PTPRT	in
ALGA0095311	17	50212488	7, 8, 9 and 10	-	-	-
ALGA0095334	17	50796444	12, 13, 14, 15 and 16	-	-	-
ASGA0102680	17	51038893	10, 11 and 12	-	-	-
ASGA0105240	17	52040070	13, 14, 15 and 16	-	TOX2	in
M1GA0025144	18	3045578	10, 11 and 12	-	-	-
MARC0010873	18	3048405	10 and 11	-	-	-
MARC0056921	18	6735781	7 and 8	-	LOC102164945 /KCNH2 AOC1 KCNH2 GIMAP2	18440 5426 10150 103

* For the HYM 13, this block is composed by markers ALGA0017362 and ASGA0013249. ** This Block appears for HYM 9 and 10.

Chapter 6

GENERAL DISCUSSION

Introduction

In genomic era, most SNP association studies have considered continuous phenotypes under Gaussian assumptions. However, reproductive traits, such as stillborn, must be characterized as discrete variable, which could potentially follow other distributions, such as Poisson. Besides, another point that deserves to be highlighted in genome wide association studies (GWAS) is the genetic dissection of complex phenotypes through candidate genes network derived from significant SNPs. In this chapter, I discuss the importance of a proper distribution analysis in GWAS. Moreover, based on post-GWAS results, the genetic architecture and environment divergences implications and their perspectives in animal breeding is discussed.

Data distribution in GWAS

In the previous chapters we have seen the importance in considering a proper distribution when working with counting data on GWAS. On this work, the Poisson model showed the best fit for stillborn trait (SB), while the Gaussian model was more suitable for the number of teats (NT) on both data sets (UFV and Topigs Norsvin population). Most studies do not consider this particular distribution of SB (Canario et al., 2006), even though the use of different discrete models, such as Poisson and binomial, can lead to more appropriate quantification of the genetic influences on this trait. Working under a hierarchical Bayesian approach, Varona and Sorensen (2010) also showed the importance in propose and compare discrete models to be fitted to stillborn data under genetic and animal breeding approach. Thus, using genomic information (i.e. under genomic approaches), the predictions of genomic estimated

breeding values (GEBVs) and marker effects estimation exploiting counting models increased our analyses accuracy.

Futhermore, even though NT is also characterized as a counting discrete variable, Gaussian distribution fitted the behavior of this trait better than Poisson. A possible reason is that Poisson distribution is asymmetric and right skewed, and the symmetry of Gaussian was more consistent with the observed distribution of the NT sample data. Another possible explanation is that the Poisson distribution assumes the mean equal to its variance, a condition that may not have been met when working with NT.

Based on these results, we conclude that distribution analyses might be useful to define the best model for counting data before performing GWAS, leading to more confidence. Even with a narrow number of studies comparing different models for discrete variables, the predictions of GEBVs and marker effects estimation exploiting counting models for them we showed increased genome wide analyses accuracy. In future studies. Concluding, we have shown that it is interesting to consider different distributions for discrete random variables, before proceeding on genomic analysis, e.g. negative binomial and generalized Poisson distributions.

Genetic architecture

The GWAS described at the present work for number of teats enabled us to observe different sets of SNPs in different populations. Inconsistence of significant markers and associated quantitative trait locus (QTL) in different populations has also been observed by Veroneze et al. (2014), where observed differences in linkage disequilibrium patterns and persistence of phase across pig lines, suggested different QTL and consequently distinct set of genes under selection. The gene-TF networks analysis in the fourth chapter, further illustrated the phenomenon observed in the

previous chapters, showing that the same trait can be under selection without selecting for the same set of genes.

According to the genetic architecture, divergences across populations might have implications. Selection acts over distinct genomic regions and depending on the population, initial selection can be on different genes and this may directly have an influence on which regions are under selection at later stages. Thereby, different genes involved in the same biological pathways may result in variations of a given trait and can be incidentally selected resulting in the same phenotypic outcome. The first implication is that distinct set of genes must be relevant for each population. Second, as distinct sets of SNPs are observed according to the population, using a single reference population on breeding programs might be a pitfall procedure.

On genome wide selection, the number of individuals in the training data set might affect the accuracy on breeding values estimates (Goddard, 2009). Thus, the use of a limited reference set is not desirable in breeding programs. One possibility was given by Chen et al., (2014) who proposed a multi-task Bayesian learning model applied for multi-population genomic prediction that consider each genomic population information. By this way, the genomic architecture divergences from each population demonstrated on our post-GWAS analyses might be useful in a combined approach increasing genomic analyses accuracy since it can indicate, due to biological roles and processes, the main genes in each candidate genomic region.

Environment divergences

Besides genomic architecture differences across populations, we demonstrated that for divergent environments, we can also note distinct set of SNPs for the same trait. With similar results, Silva et al. (2014) observed also variation in the significant SNP set for total number of piglets born in a commercial line according to environment

variations resulting in different QTLs. Moreover, Machado et al., (2010) in a QTL study for tick resistance/susceptibility in a bovine F2 population found that depending on the tick evaluation season (dry and rainy), different sets of genes could be involved in the tick resistance mechanism.

In this study, the gene-TF networks analysis further illustrate the phenomenon observed in these previous studies, showing that the same trait can be under selection without selecting for the same sets of genes. Through the environment divergences point of view, selection can act on many genes depending on specifics scenarios. Thus, different genes involved in distinct biological pathways may illustrate the physiological requirement for the trait according to environment. Therefore, different components of animal physiology (e.g. reproductive and endocrine systems) may be important for determining the final trait, and different genes are therefore important and contribute to the observed genetic differences.

Data analyses from different environments are valuable for animal breeding as they provide additional insights into the genes controlling traits of interest under distinct scenarios. International breeding companies might consider the gene sets specificity in order to perform proper association studies according environments. Thus, strategies can be taken in view of the breeding interests. Moreover, these findings demonstrate the need for a biological understanding of complex traits in genomic studies across environments.

Concluding remarks

The present Post-GWAS study provided a rich information resource about genes identified using genome wide association approaches for reproductive traits. The distribution analyses in genomic models highlighted the importance to consider

counting models. Moreover, different sets of relevant SNPs and QTL blocks across and within the studies were identified leading to the possibility of different set of genes playing biological roles related to a single complex trait. Thereby, we highlighted the genomic diversity across population/environments to be observed in breeding programs in such a way that population/environments specific reference populations might be considered in genomic analyses. Based on these results, we demonstrated the importance of post-GWAS analyses to increase the biological understanding of relevant genes for complex traits.

REFERENCES

Canario L, Cantoni E, Le Bihan E, Caritez JC, Billon Y, Bidanel JP, et al.: **Between-breed variability of stillbirth and its relationship with sow and piglet characteristics**. *J Anim Sci* 2006, **84**:3185–3196.

Chen L, Li C, Miller S, Schenkel F: **Multi-population genomic prediction using a multi-task Bayesian learning model.** *BMC genetics* 2014, **15**(1):53.

Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136**(2)**:**245-257.

Machado MA, Azevedo AL, Teodoro RL, Pires MA, Peixoto MG, de Freitas C, et al.: **Genome wide scan for quantitative trait loci affecting tick resistance in cattle (*Bos taurus* x *Bos indicus*)**. *BMC Genomics* 2010, **11**:280.

Silva FF, Mulder HA, Knol EF, Lopes MS, Guimarães SE, Lopes PS, et al.: **Sire evaluation for total number born in pigs using a genomic reaction norms approach**. *J Anim Sci* 2014, **92**(9):3825-3834.12.

Varona L, Sorensen D: **A genetic analysis of mortality in pigs**. *Genetics* 2010, **184**:277–284.

Veroneze R, Bastiaansen J, Knol EF, Guimarães S, Silva FF, Harlizius B, et al.: **Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations**. *BMC Genet* 2014, **15**(1):126.