J Appl Genetics (2015) 56:123–132
DOI 10.1007/s13353-014-0240-y

ANIMAL GENETICS • ORIGINAL PAPER

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

L. L. Verardo · F. F. Silva · L. Varona · M. D. V. Resende · J. W. M. Bastiaansen · P. S. Lopes · S. E. F. Guimarães

Received: 7 March 2014 / Revised: 27 May 2014 / Accepted: 23 July 2014 / Published online: 8 August 2014

Abstract The genetic improvement of reproductive traits such as the number of teats is essential to the success of the pig industry. As opposite to most SNP association studies that consider continuous phenotypes under Gaussian assumptions, this trait is characterized as a 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 Counting data · Genes · Reproductive traits · SNP association

Electronic supplementary material The online version of this article (doi:10.1007/s13353-014-0240-y) contains supplementary material, which is available to authorized users.

L. L. Verardo (✉) · F. F. Silva · P. S. Lopes · S. E. F. Guimarães
Department of Animal Science, Universidade Federal de Viçosa - UFV, Viçosa, MG, Brazil
e-mail: lucas_verardo@yahoo.com.br

L. Varona
Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, Zaragoza, Spain

M. D. V. Resende
Embrapa Florestas, Colombo, PR, Brazil

J. W. M. Bastiaansen
Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands

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 trait 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. 2011a), 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. 2011a) 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

A hierarchical Bayesian multiple regression model considering two different distributions for the data, Gaussian and Poisson, was proposed. 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, 1, and 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, $\boldsymbol{\beta}=[\beta_1, \beta_2, \ldots, \beta_{237}]'$, $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, 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 (λ_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, 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. A total of 100,000 iterations with a burn-in period and sampling interval (thin) of 50,000 and five iterations, respectively, were considered. 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 $D_G^{(j)} = -2\log\left(\prod_{i=1}^{345} P\left(y_i|\theta^{(j)}\right)\right)$, in which $\prod_{i=1}^{345} P\left(y_i|\theta^{(j)}\right)$ is the value for the likelihood function considering the set of parameter estimates ($\theta^{(j)}$) at each MCMC iteration j. Similarly, for the Poisson model, the values came from $D_P^{(j)} = -2\log\left(\prod_{i=1}^{345} P\left(y_i|\lambda_i^{(j)}\right)\right)$, being $\lambda_i^{(j)}$ the estimate of Poisson mean, i.e., the exponential of the latent variable log(λ_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 this distribution was obtained, it was possible to propose a hypothesis test based on highest posterior density (HPD) 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 $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, was also included in models (1) and (2) to point out 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**), $\mathbf{Z}^* = \mathbf{ZL}$, 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 ($\widehat{\mathbf{u}}^*$) must be used to obtain the original vector of breeding values, which is given by σ^2_e~ IGamma(a, 1/b). 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 the best model is identified, the significance of SNP effects can also 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 it 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 then identify 19 significant SNPs using a hypothesis test based on 95 % HPD interval for their effects (Table 1).

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).

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).

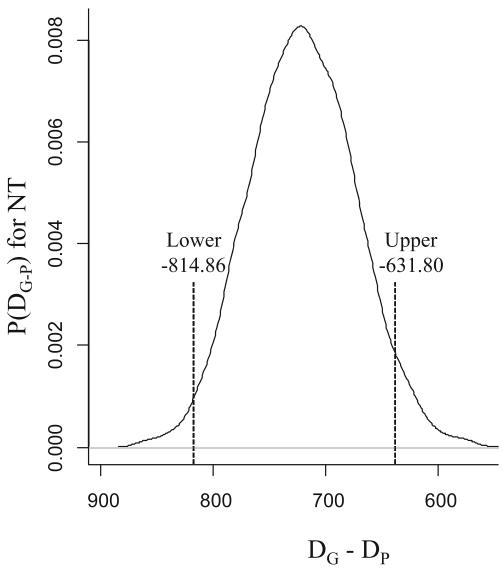

Fig. 1 Distribution of deviance difference, $P(D_{G-P})$, between Gaussian and Poisson GWAS models fitted to number of teats (NT) in pigs. D_G and D_P are the estimated Gaussian and Poisson models deviance, respectively

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

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

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. (2011b), 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 also found 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.

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

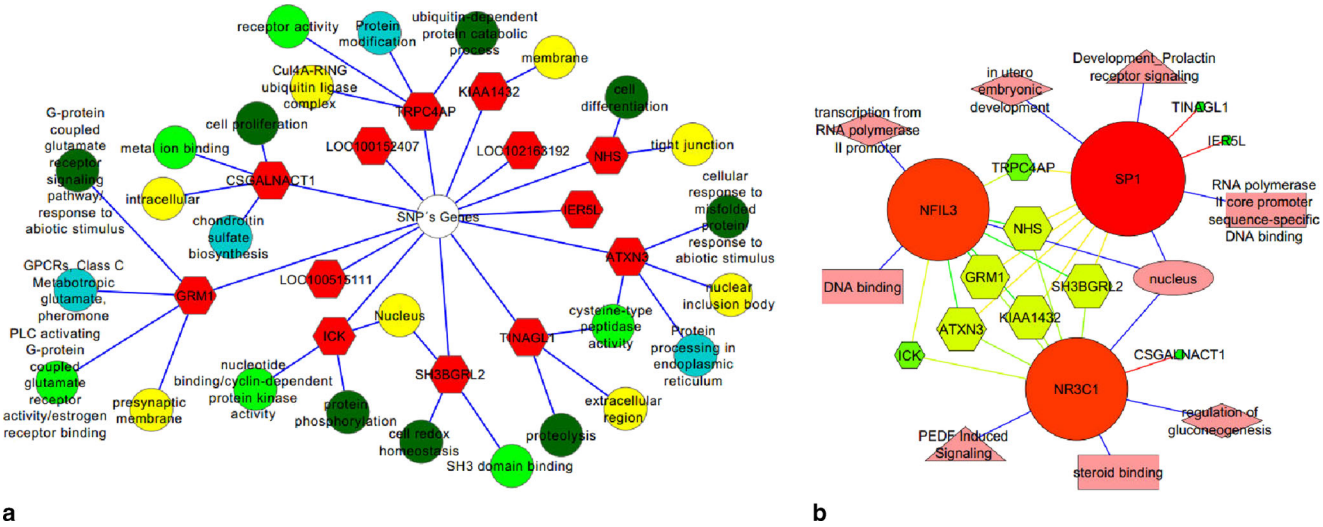

Fig. 2 Functional networks and their interactions for number of teats trait. **a** The relationships between 13 genes (in *octagon red*) and their related subnets: pathway (in *blue*), biological process (in *dark green*), molecular function (in *light green*), and cellular component (in *yellow*). **b** 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*)

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. Cepica et 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 (i.e., 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 be involved in signal transduction, cytoskeleton rearrangements, membrane trafficking, and other key cellular processes (Cesareni et al. 2002). Here this gene has a polymorphism in it sequence related to number of teats trait and it shars 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 also identified genes sharing the same biological process as a 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 1 receptors. 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). ATXN3 shares this biological process with this gene. This gene is also known as Machado-Joseph disease (MJD) and encodes for Ataxin-3 protein which might a play role 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 its expression altered when induced by stradiol, diethylstilbestrol, and octyl-phenol in the uterus of immature rats (Hong et al. 2006).

The ATXN3 also shares 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 1 and 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 with 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 1 which encodes a glucocorticoid receptor (GR). Studies suggest that GR regulates 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 Abdelrahim 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 evidence 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 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 to develop the 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 for 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 research works generalizations of this model can be proposed in order to point out the 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 of a large number of genes that are regulated by several transcription factors, therefore many of them still need to be identified.

Acknowledgments 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.

References

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. Animal 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? Fed Eur Biochem Soc 513:38–44

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 region-specific microarrays reveal differential expression of positional candidate genes of signalingpathways associated withthe liability forthe inverted teat defect. Anim Genet 44(2):139–148

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

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. Genetics 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 JNN, PrayagaKC BW, Hawken RJ (2010) Associationweight matrix for the genetic dissection of puberty in beef cattle. Proc Natl Acad Sci U S A 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

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):486–495

Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182:343–353

Harville DA, Callanan TP (1989) Computational aspects of likelihood-based inference for variance components. In: Gianola D, Hammond K (eds) Advances in statistical methods for genetic improvement of livestock, edn. Springer, Berlin, pp 136–176

Hidalgo AM, Lopes PS, Paixão DM, Silva FF, BastiaansenJWM PSR, 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. Reprod 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 Bioinforma 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

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

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

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. PLoS One 8(7):e69851

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

Onteru SK, Fan B, Du Z-Q, Garrick DJ, Stalder KJ, Rothschild MF (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, Fortes MRS (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, Rempel LA, 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 Neurol 322:121–135

Silva KM, Knol EF, Merks JWM, Guimarães SEF, Bastiaansen JWM, van Arendonk JAM, Lopes PS (2011a) 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 (2011b) Three-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. Livest Sci 142(1):210–215

Smith BJ (2007) Boa: an R package for MCMC output convergence assessment and posterior inference. J Stat Softw 21(11):1–37

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. Algorithm 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, Sorensen D (2010) A genetic analysis of mortality in pigs. Genetics 184:277

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

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):376–384

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. Med 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. Bioinformatics 28(24):3326–3328