UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

GENOME-WIDE SCANS TO UNCOVER LOCI UNDERLYING BOVINE REPRODUCTIVE BIOLOGY

Yuri Tani Utsunomiya
Médico Veterinário

2013

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

GENOME-WIDE SCANS TO UNCOVER LOCI UNDERLYING BOVINE REPRODUCTIVE BIOLOGY

Yuri Tani Utsunomiya

Orientador: Prof. Adj. José Fernando Garcia

Co-orientador: Prof. Dr. Johann Sölkner

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para obtenção do título de Mestre em Medicina Veterinária (Reprodução Animal).

2013

Utsunomiya, Yuri Tani
U89g Genome-wide scans to uncover loci underlying bovine reproductive biology / Yuri Tani Utsunomiya. – – Jaboticabal, 2013
xii, 106 p. : il. ; 28 cm

Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2013
Orientador: José Fernando Garcia
Banca examinadora: Joaquim Mansano Garcia, Adriana Santana do Carmo
Bibliografia

1. *Bos indicus*. 2. Perímetro escrotal. 3. Fertilidade. 4. Peso ao nascer. 5. *PLAG1*. 6. SNP. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 619:636.082:636.2

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação – Serviço Técnico de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

CAMPUS DE JABOTICABAL

FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS DE JABOTICABAL

CERTIFICADO DE APROVAÇÃO

TÍTULO: GENOME-WIDE SCANS TO UNCOVER LOCI UNDERLYING BOVINE REPRODUCTIVE BIOLOGY

AUTOR: YURI TANI UTSUNOMIYA

ORIENTADOR: Prof. Dr. JOSE FERNANDO GARCIA
CO-ORIENTADOR: Prof. Dr. JOHANN SOLKNER

Aprovado como parte das exigências para obtenção do Título de MESTRE EM MEDICINA VETERINÁRIA , Área: REPRODUÇÃO ANIMAL, pela Comissão Examinadora:

Prof. Dr. JOSE FERNANDO GARCIA
Departamento de Apoio, Produçao e Saude Animal / Faculdade de Medicina Veterinária de Araçatuba

Prof. Dr. JOAQUIM MANSANO GARCIA
Departamento de Medicina Veterinária Preventiva e Reprodução Animal / Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal

Adriana Santana do Carmo
Profa. Dra. ADRIANA SANTANA DO CARMO
Deoxi Biotecnologia Ltda / Araçatuba/SP

Data da realização: 16 de dezembro de 2013.

ABOUT THE AUTHOR

YURI TANI UTSUNOMIYA – middle child of Josiane Cristina Harth Scanzani and Takashi Utsunomiya, born March 21st 1988 in Campinas-SP, Brazil. Utsunomiya graduated at the Faculty of Veterinary Medicine of Araçatuba (UNESP) with a Médico Veterinário (Bachelor of Veterinary Medicine) degree in 2010, when he undertook an internship in Molecular Genetics and Bioinformatics at Università Cattolica del Sacro Cuore (Piacenza, Italy) under the supervision of Prof. Paolo Ajmone-Marsan. He started his Master of Science (MSc.) studies under the Graduate Program in Veterinary Medicine of FCAV-UNESP, Jaboticabal campus, in August 2011, focusing in the field of Animal Reproduction. From April to July 2012, Utsunomiya was a visiting junior researcher supported by the European Science Foundation and the Advances in Farm Animal Genomic Resources project at the University of Natural Resources and Life Sciences (Vienna, Austria), under the supervision of Prof. Johann Sölkner, where part of his MSc. studies were conducted. As a graduate student at UNESP and MSc. fellow with Prof. José Fernando Garcia at FCAV-UNESP, Utsunomiya has been supported by the São Paulo Research Foundation (FAPESP), and is co-author of 8 peer-reviewed articles, 23 abstracts, and one open source software.

“Ubuntu does not mean that people should not enrich themselves. The question therefore is: Are you going to do so in order to enable the community around you to be able to improve?”

Nelson Mandela

ACKNOWLEDGMENTS

To my parents, Josiane Cristina Harth Scanzani and Takashi Utsunomiya, for being my motivators and my balance.

To my family, to whom I credit all my personal achievements.

To Rafaela Beatriz Pintor Torrecilha, for being my best friend. Although you are biased, I love your advices.

To Prof. José Fernando Garcia, who credited to me trust and showed me the pathway of science. I will have in mind our first conversation in all my professional decisions.

To Dr. Adriana Santana do Carmo, who as a pioneer paved the way so others could hit the road, travel, and dream.

To Prof. Cáris Maroni Nunes and all my dear friends from the Animal Biochemistry and Molecular Biology Laboratory, for all support, patience and friendship in the last six years.

To Prof. Johann Sölkner and my new friends from the University of Natural Resources and Life Sciences Vienna, for opening the doors of their laboratory and granting me an extraordinary experience.

To the Zebu Genome Consortium and Conexão Delta G, for providing data and technical and scientific counseling to make this work possible.

To the São Paulo Research Foundation (FAPESP) and the European Science Foundation and the Advances in Farm Animal Genomic Resources project, for financially supporting this research.

And finally, to the Faculdade de Ciências Agrárias e Veterinárias, FCAV-UNESP, Jaboticabal campus, for providing me technical trainning and scientific ambience.

Yuri Tani Utsunomiya

Jaboticabal, November 2013.

SUMMARY

VARREDURAS GENÔMICAS PARA A DETECÇÃO DE LOCI ENVOLVIDOS NA BIOLOGIA REPRODUTIVA DE BOVINOS

RESUMO – O desempenho reprodutivo dos animais tem um grande impacto sobre a indústria da carne bovina. A caracterização de regiões genômicas que afetam a fertilidade dos animais pode contribuir para a identificação de marcadores preditivos de desempenho reprodutivo e desvendar os mecanismos moleculares envolvidos em aspectos complexos da biologia reprodutiva dos bovinos. Nos dois primeiros estudos relatados, os genomas de touros da raça Nelore (*Bos indicus*) foram examinados em busca de loci que explicam variação nas características peso ao nascer (PN) e perímetro escrotal (PE), utilizando dados de mais de 777.000 marcadores do tipo polimorfismo de sítio único (*single nucleotide polymorphism* - SNP). Um segmento do cromossomo 14, o qual engloba o gene ortólogo *PLAG1* que afeta estatura em humanos, foi encontrado tanto em PN quanto em PE. Este locus possui efeitos pleiotrópicos sobre características reprodutivas e de tamanho corporal em bovinos, e representa um ponto de partida para a dissecção da genética da fertilidade bovina. Em outro estudo, um teste estatístico composto foi desenvolvido e aplicado na busca de evidências de assinaturas de seleção no genoma de raças bovinas de leite e de corte. Padrões de variação genética que podem ter sido moldadas pela seleção humana foram detectados no genoma de quatro diferentes raças bovinas (Angus, Pardo Suíço, Gir e Nelore). O estudo indica o gene Cornichon 3 (*CNIH3*) como um forte candidato, que pode estar envolvido na regulação do pico pré-ovulatório do hormônio luteinizante na raça Pardo-Suíço. Embora estes resultados apenas toquem a superfície dos mecanismos moleculares por trás da reprodução dos bovinos, os loci aqui identificados abrigam novos e conhecidos genes candidatos que afetam a fertilidade da espécie, e oferecem novas perspectivas sobre aspectos complexos de sua biologia reprodutiva.

Palavras-chave: *Bos indicus*, Perímetro escrotal, Fertilidade, Peso ao nascer, *PLAG1*, SNP

GENOME-WIDE SCANS TO UNCOVER LOCI UNDERLYING BOVINE REPRODUCTIVE BIOLOGY

ABSTRACT – Reproductive performance has a high impact on the beef cattle industry. The characterization of genomic regions affecting fertility can contribute to the identification of diagnostic markers for reproductive performance and uncover molecular mechanisms underlying complex aspects of bovine reproductive biology. In the first two reported studies, the genomes of progeny-tested Nellore bulls (*Bos indicus*) were scanned for loci explaining variance in birth weight (BW) and scrotal circunferemce (SC), using data containing over 777,000 single nucleotide polymorphism (SNP) markers. Among the identified loci, a chromosome segment located on autosome 14, encompassing the orthologous human stature gene pleiomorphic adenoma 1 (*PLAG1*), was found to affect both BW and SC. This locus has been found to have pleiotropic effects on reproduction and body size traits in cattle, and represents a starting point to the dissection of the complex inheritance of bovine fertility. In a separate study, a composite statistical test was developed and applied to scan dairy and beef cattle genomes for evidences of natural and artificial selection signatures. Patterns of genetic variation that may have been shaped by human-driven selection were detected in the genomes of four different cattle breeds (Angus, Brown Swiss, Gyr and Nellore). The study pointed to the Cornichon homolog 3 gene (*CNIH3*) as a strong candidate involved in the regulation of pre-ovulatory luteinizing hormone surge in Brown Swiss. Although these findings only scratch the surface of the molecular mechanisms underlying bovine reproduction, the loci identified here harbor known and novel functional candidate genes affecting fertility in cattle and offer new insights on complex aspects of bovine reproductive biology.

Keywords: *Bos indicus*, Scrotal circumference, Fertility, Birth weight, *PLAG1*, SNP

LIST OF ABBREVIATIONS AND SYMBOLS

A	Adenine
Ala	Alanine
AMPA	Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ANG	Angus
Asn	Asparagine
BSW	Brown Swiss
BTA	*Bos taurus*
BW	Birth weight
C15ORF55	Nuclear protein in testis (*NUTM1*) gene
CES4A	Carboxylesterase 4A (*CES6*, *CES8*, Hydrolase A) gene
CHCHD7	Coiled-coil-helix-coiled-coil-helix domain containing 7 gene
CI	Confidence interval
CMS	Composite of multiple signals
CR_{IND}	Individual call rate
CR_{SNP}	SNP call rate
DAVID	Database for annotation, visualization and integrated discovery
dEBV	Deregressed estimated breeding value
DNA	deoxyribonucleic acid
e.g.	*Exempli gratia* (latim for “for example”)
EBV	Estimated breeding value
EHH	Extended haplotype homozygosity
FASTA	Fast association score test-based analysis
FSIP1	Fibrous sheath-interacting protein 1 gene
G	Guanine
GC	Genomic control
GenCall	Genotype call
GnRH	Gonadotropin-releasing hormone
GRIA	Glutamate Receptor, Ionotropic, AMPA
GWAS	Genome-wide association study
GYR	Gyr

HDL	High density lipoprotein
HSA	*Homo sapiens*
HWE	Hardy-Weinberg equilibrium
i.e.	*Id est* (latim for “that is” or “in other words”)
IBD	Identical by descent; identity by descent
IBS	Identity by state; identical by state
IGF1	Insulin-like growth factor 1
IMPAD1	Inositol monophosphatase domain-containing protein 1 gene
kb	Kilobase (10^3 nucleotides)
kg	Kilograms
KIT	Mast/stem cell growth factor receptor gene
LCORL	Ligand dependent nuclear receptor corepressor-like gene
LD	Linkage disequilibrium
LH	Luteinizing hormone
LYN	Tyrosine-proteinkinase Lyn gene
MAD2	mitotic arrest deficient-like gene
MAF	Minor allele frequency
MAGEL2	Melanoma antigen family L 2 gene
Mb	Megabase (10^6 nucleotides)
MC1R	melanocortin 1 receptor gene
MMU	*Mus musculus*
MOS	V-mos Moloney murine sarcoma viral oncogene homolog gene
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
NCAPG	Non-SMC condensin I complex, subunit G gene
NEL	Nellore
PAR	XY pseudo-autosomal region
PCoA	Principal Coordinates Analysis (Classical Multidimensional Scalling)
PDE5A	Phosphodiesterase 5A gene
PENK	Proenkephalin-A gene
PLAG1	Pleiomorphic adenoma 1 gene

QC	Quality control
Q-Q	Quantile-quantile
QTL	Quantitative trait locus
QTN	Quantitative trait nucleotide
RDHE2	Epidermal retinol dehydrogenase 2 (*SDR16C5*) gene
RNA	Ribonucleic acid
RPS20	40S ribosomal protein S20 gene
rRNA	Ribosomal RNA
SC	Scrotal circumference
SC_A	Scrotal circumference corrected for age at yearling
SC_{AW}	Scrotal circumference corrected for age and weight at yearling
SDR16C5	Epidermal retinol dehydrogenase 2 (RDHE2) gene
SDR16C6	Short-chain dehydrogenase / reductase family 16C gene
Ser	Serine
SH3RF2	SH3 domain containing ring finger 2 gene
SNAI2	Snail family zinc finger 2 gene
SNP	Single nucleotide polymorphism
SP4	Sp4 transcription factor gene
SSC	*Sus scrofa*
ST6GALNAC5	ST6(Alpha-N-Acetyl-Neuraminyl-2,3-Beta-Galactosyl-1,3)-N-Acetylgalactosaminide Alpha-2,6-Sialyltransferase 5 gene
TGS1	Trimethylguanosine synthase 1 gene
Thr	Threonine
U6	U6 spliceosomal RNA gene
XKR4	Kell blood group complex subunit-related family, member 4 (*KIAA1889*) gene

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 - General considerations

1. Introduction

Descendants from the extinct aurochs (*Bos primigenius*), the humpless taurine (*Bos taurus*) and the humped indicine cattle (zebu, *Bos indicus*) were domesticated some 8,000 years B.C. in Southwestern and Southern Asia, respectively, and spread through the world accompanying our species due to the expansion of agriculture (LOFTUS et al., 1994; BRUFORD et al., 2003). Within their recent evolutionary history, humans and cattle colonized the world together (AJMONE-MARSAN et al., 2010), and today, the ability of the bovine species to convert low-quality forage into meat, milk and draft power is of direct importance to the livelihood of over 6.6 billion people (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al., 2009).

By 2050, global meat production will need to be doubled relative to the current production levels in order to feed over 9.2 billion people, and these figures will need to be achieved under strict socio-environmental sustainable guidelines (FAO, 2010). In this scenario, the cattle industry needs to undergo deep innovation in order to increase production efficiency.

One way to meet growing demands under such constraints is by selecting animals with above-average productive performance as parents of the next generation. This has been traditionally done by the industry through progeny testing, assessment of sires estimated breeding values (EBV) and use of assisted reproductive technologies (GARCIA et al., 2013). However, measuring and evaluating progeny performance has been a major challenge due to its high cost and time consuming nature, and more efficient alternatives must be sought.

The first critical part in selecting superior animals is the establishment of a list of cost-effective measurable traits that have a high impact in the activity. According to Garrick (2011), the major economically important traits in beef cattle are those related to reproductive performance, growth rate, and survival. The reproductive performance of animals is of particular importance as it affects generation intervals, the rate of genetic change, and the amount of product that can be sent to the market

(VAN MELIS et al., 2010), and age at puberty, age at first conception, duration of post-partum anoestrus and total lifetime productivity are the main factors influencing reproductive performance in cattle (BURNS et al., 2010).

Ideally, reproductive traits for selection would be moderately heritable, measured early in life and correlated with future mating performance (FORTES et al., 2012). However, traits considered indicative of reproductive performance generally exhibit complex inheritance (i.e., the genetic variance accounts for only a fraction of the total trait variance, and many small effects genetic loci contribute to genetic variation), and are expressed late in the life of an animal (CAMMACK et al., 2009). Hence, the molecular dissection of the complex genetic architecture underlying fertility and correlated traits may be of great importance to the identification of predictive markers for the improvement of selection for reproductive performance.

In recent years, the release of reference genome assemblies, together with initiatives for genome re-sequencing, has enabled the discovery of millions of single nucleotide polymorphism (SNP) markers across populations in several species of animals, including cattle (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al., 2009; THE BOVINE HAPMAP CONSORTIUM et al., 2009). These genetic markers are bi-allelic loci that are highly abundant and tightly distributed across the genome (KIM & MISRA, 2007). As SNPs that are located nearby in a chromosome region exhibit high correlations (i.e., linkage disequilibrium), one can assay few tens of thousands of these genetic markers and use them as surrogates for unobserved variants across the genome (BOHMANOVA et al., 2010). The development of high-throughput DNA microarray technologies has recently allowed for the rise of low cost large-scale genotyping, and it is now possible to profile hundreds of thousands of SNP markers in a single bovine DNA sample (MATUKUMALLI et al., 2009).

Genomic selection, the concept of using high-density SNP information to predict EBVs, was first introduced by Meuwissen et al. (2001), and is the main present date driver of the generation of large amount of cattle genomic data. The approach consists in estimating the relative breeding values of individual chromosome fragments, and then summing up the values of all inherited chromosome fragments in a selection candidate in order to obtain a molecular

estimate of its genetic merit. Genomic selection has been propagated as a “paradigm shifting” innovation in the sector in recent years, and the use of genomic information to predict breeding values has revolutionized the dairy cattle industry and is now being implemented in beef cattle (GARRICK et al., 2011; GARCIA et al., 2013).

Although genomic selection is currently the main stream application of genomics in the beef cattle industry, the generation of large amount of genomic data has the potential to take us beyond predicting genetic merit. It is becoming clear that linking phenotypes to gene function can generate invaluable knowledge to develop new technologies. By using SNP data, it is now possible to scan whole cattle genomes, and uncover loci explaining variance in traits of interest (BUSH & MOORE, 2012), or even seek genomic regions where past selection has taken place (OLEKSYK et al., 2010). As this wealth of data offers a new opportunity to shed light into complex aspects of bovine reproductive biology, the objective of this Master of Science dissertation was to perform genome-wide scans to identify putative loci underlying reproductive performance in cattle.

In Chapter 2, the genomes of over 600 Nellore sires were scanned in the search of loci affecting birth weight. Although the beef industry pursues animals with heavy carcasses, selecting animals with high weights at birth often results in decreased reproductive performance due to increased rates of dystocia (COOK et al. 1993) and perinatal mortality (JOHANSON & BERGER, 2003). Identifying loci that individually affect weight and fertility traits, as well as variants with pleiotropic effects, may be of help to balance these conflicting selection goals. A strong signal was found on chromosome 14, surrounded by several genes previously demonstrated to affect stature in cattle and humans (FORTES et al., 2012). This genomic region includes the pleomorphic adenoma 1 gene (*PLAG1*), which has been recently found to have pleiotropic effects on fertility and body size traits in cattle (FORTES et al., 2013). This study was published in *BMC Genetics* in June 2013 (UTSUNOMIYA et al., 2013a).

In Chapter 3, a genome-wide mapping for chromosome segments explaining differences in scrotal circumference at yearling in Nellore cattle is reported. Putative loci affecting the trait were identified on chromosomes 4, 6, 7, 10, 14, 18 and 21. Interestingly, the locus encompassing *PLAG1* was again detected, replicating the

evidence that this gene has a key role in fertility and body size in cattle. Novel candidate genes that affect growth and testicular size in other animal models were also identified, including *SP4*, *MAGEL2*, *SH3RF2*, *PDE5A* and *SNAI2*. This study has been submitted for publication, and is currently under peer-review.

Finally, in Chapter 4, a simple approach for combining different ways to scan genomes for evidence of signatures of natural and artificial selection is described and applied to dairy and beef cattle. Patterns of genetic variation that may have been shaped by human-driven selection were detected in the genomes of four different cattle breeds (Angus, Brown Swiss, Gyr and Nellore), and represent important resources for characterizing genome regions that affect economically important traits. In particular, the most significant SNP identified is intronic to the Cornichon homolog 3 gene (*CNIH3*), and may be involved in the regulation of pre-ovulatory luteinizing hormone surge. This study was published in *PLoS ONE* in May 2013 (UTSUNOMIYA et al., 2013b).

2. References

Ajmone-Marsan, P.; Garcia, J. F.; Lenstra, J. A.; Globaldiv Consortium. On the origin of cattle: How aurochs became cattle and colonized the world. **Evolutionary Anthropology**, v. 19, p. 148-157, 2010.

Bohmanova, J.; Sargolzaei, M.; Schenkel, F. S. Characteristics of linkage disequilibrium in North American Holsteins. **BMC Genomics**, 11:421, 2010.

Bruford, M. W.; Bradley, D. G.; Luikart G. Genetic analysis reveals complexity of livestock domestication. **Nature Reviews Genetics**, v. 4, p. 900-910, 2003.

Burns, B. M.; Fordyce, G.; Holroyd, R. G. A review of factors that impact on the capacity of beef cattle females to conceive, maintain a pregnancy and wean a calf-implications for reproductive efficiency in northern Australia. **Animal Reproduction Science**, v. 122, p. 1-22, 2010.

Bush, W. S.; Moore, J. H. Genome-wide association studies. **PLoS Computional Biolology**, 8:e1002822, 2012.

Cammack, K. M., Thomas, M.G, Enns, R.M. Review: Reproductive traits and their heritabilities in beef cattle. **The Professional Animal Scientist**, v. 25, p. 517-528. 2009.

Cook, B. R.; Tess, M. W.; Kress, D. D. Effects of selection strategies using heifer pelvic area and sire birth weight expected progeny difference on dystocia in first-calf heifers. **Journal of Animal Science**, v. 71, p. 602-607, 1993.

FAO - Food and Agriculture Organization of the United Nations. The state of food and agriculture 2009: livestock in the balance. **FAO/Eletronic Publishing Policy and Support Branch**, 166p, 2010.

Fortes, M. R. S.; Kemper, K.; Sasazaki, S.; Reverter, A.; Pryce, J. E.; Barendse, W.; Bunch, R.; McCulloch, R.; Harrison, B.; Bolormaa, S.; Zhang, Y. D.; Hawken, R. J.; Goddard, M. E.; Lehnert, S. A. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. **Animal Genetics**, in press, 2013.

Fortes, M. R. S.; Reverter, A.; Hawken, R. J.; Bolormaa, S.; Lehnert, S. A. Candidate genes associated with testicular development, sperm quality, and hormone levels of inhibin, luteinizing hormone, and insulin-like growth factor 1 in Brahman bulls. **Biology of Reproduction**, 87:58, 2012.

Garcia, J. F.; Alonso R. V.; Utsunomiya, Y. T.; Carmo, A. S. Genomic selection and assisted reproduction technologies to foster cattle breeding. **Animal Reproduction**, v.10, n.3, p.297-301, 2013.

Garrick D. J. The nature, scope and impact of genomic prediction in beef cattle in the United States. **Genetics Selection Evolution**, 43:17, 2011.

Johanson, J. M.; Berger, P. J. Birth weight as a predictor of calving ease and perinatal mortality in Holstein cattle. **Journal of Dairy Science**, v. 86, n. 11, 3745-3755, 2003.

Kim S., Misra A. SNP genotyping: Technologies and biomedical applications. **Annual Review of Biomedical Engineering**, v. 9, p. 289-320, 2007.

Loftus, R. T.; MacHugh, D. E.; Bradley, D. G.; Sharp, P. M.; Cunningham, P. Evidence for two independent domestications of cattle. **Proceedings of the National Academy of Science U. S. A.**, v. 91, p. 2757-2761, 1994.

Matukumalli, L. K.; Lawley, C. T.; Schnabel, R. D.; Taylor, J. F.; Allan, M. F.; Heaton, M. P.; O'Connell, J.; Moore, S. S.; Smith, T. P.; Sonstegard, T. S.; Van Tassell, C. P. Development and characterization of a high density SNP genotyping assay for cattle. **PLoS One**, 4:e5350, 2009.

Meuwissen, T. H. E.; Hayes, B.; Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2011.

Oleksyk, T. K.; Smith, M. W.; O'Brien, S. J. Genome-wide scans for footprints of natural selection. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 365, p. 185-205, 2010.

The Bovine Genome Sequencing and Analysis Consortium; Elsik, C. G.; Tellam, R. L.; Worley, K. C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, v. 324, n. 5926, p. 522-528, 2009.

The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. **Science**, v. 324, n. 5926, p. 528-532, 2009.

Utsunomiya, Y. T.; Carmo, A. S.; Carvalheiro, R.; Neves, H. H.; Matos, M. C.; Zavarez, L. B.; Pérez O'Brien, A. M. ; Sölkner, J.; McEwan, J. C.; Cole, J. B.; Van Tassell, C. P.; Schenkel, F. S.; da Silva, M. V. G. B.; Porto-Neto, L. R.; Sonstegard, T. S.; Garcia, J. F. Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. **BMC Genetics**, 14:52, 2013a.

Utsunomiya, Y. T.; Pérez O'Brien, A. M.; Sonstegard, T. S.; Van Tassell, C. P.; Carmo, A. S.; Mészarós, G.; Sölkner, J.; Garcia J. F. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. **PLoS ONE**, 8:e64280, 2013b.

Van Melis, M. H.; Eler, J. P.; Rosa, G. J.; Ferraz, J. B.; Figueiredo, L. G.; Mattos, E. C.; Oliveira, H. N. Additive genetic relationships between scrotal circumference, heifer pregnancy, and stayability in Nellore cattle. **Journal of Animal Science**, v. 88, p. 3809-3813, 2010.

CHAPTER 2 - Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height

Utsunomiya, Y. T.; Carmo, A. S.; Carvalheiro, R; Neves, H. H. R.; Matos, M. C.; Zavarez, L. B.; Pérez O'Brien, A. M.; Sölkner, J.; McEwan, J. C.; Cole, J. B.; Van Tassell, C. P.; Schenkel, F. S.; da Silva, M. V. G. B.; Porto-Neto, L. R.; Sonstegard, T. S.; Garcia J. F.

BMC Genetics 14:52, 2013
DOI: 10.1186/1471-2156-14-52

1. Abstract

Birth weight (BW) is an economically important trait in beef cattle, and is associated with growth- and stature-related traits and calving difficulty. One region of the cattle genome, located on *Bos taurus* chromosome 14 (BTA14), has been previously shown to be associated with stature by multiple independent studies, and contains orthologous genes affecting human height. A genome-wide association study (GWAS) for BW in Brazilian Nellore cattle (*Bos indicus*) was performed using estimated breeding values (EBVs) of 654 progeny-tested bulls genotyped for over 777,000 single nucleotide polymorphisms (SNPs). The most significant SNP (rs133012258, $P_{GC} = 1.34 \times 10^{-9}$), located at BTA14:25376827, explained 4.62% of the variance in BW EBVs. The surrounding 1 Mb region presented high identity with human, pig and mouse autosomes 8, 4 and 4, respectively, and contains the orthologous height genes *PLAG1*, *CHCHD7*, *MOS*, *RPS20*, *LYN*, *RDHE2* (*SDR16C5*) and *PENK*. The region also overlapped 28 quantitative trait loci (QTLs) previously reported in literature by linkage mapping studies in cattle, including QTLs for birth weight, mature height, carcass weight, stature, pre-weaning average daily gain, calving ease, and gestation length. This study presents the first GWAS applying a high-density SNP panel to identify putative chromosome regions affecting birth weight in Nellore cattle. These results suggest that the QTLs on BTA14 associated with body size in taurine cattle (*Bos taurus*) also affect birth weight and size in zebu cattle (*Bos indicus*).

Keywords: GWAS, Birth weight, *Bos indicus*, Nellore cattle, Stature

2. Introduction

Birth weight (BW) is an economically important trait in beef cattle, and is usually the first characteristic measured in a calf. Birth weight is associated with growth-related traits (BOLIGON et al., 2009), mature size (MEYER, 1995) and carcass weight, thus being a valuable production indicator, as well as a selection criterion to improve calving ease (BOURDON & BRINKS, 1982; ERIKSSON et al., 2004).

Despite the beef industry's pursuit of animals with rapid growth, yielding heavier carcasses, selection for these objectives needs to be properly balanced against selection for reproductive traits, which have great economic importance in beef cattle production systems (PHOCAS et al., 1998; GUTIÉRREZ et al., 2007). While low estimated breeding values (EBVs) for BW are associated with reduced calf viability (ERIKSSON et al., 2004) and lower growth rates (BOURDON & BRINKS, 1982; COOK et al. 1993), the use of sires with high EBVs for BW on dams with small pelvic size may result in higher rates of dystocia (COOK et al. 1993) and increased perinatal mortality (JOHANSON & BERGER, 2003). These antagonisms result from the strong association of birth weight with the body size of the calf, i.e., with the stature of the animal (MEYER, 2009). Calving difficulties can result from a mismatch between pelvic opening and calf size (GUTIÉRREZ et al., 2007). This relationship between BW with reproductive and growth/size traits highlights the importance of understanding the underlying genetic architecture of BW.

Birth weight exhibits sufficient variability and heritability in the Nellore breed (*Bos indicus*), with an average of 29.8 ± 2.7 kg (NOBRE et al., 2003) and estimated heritability between 0.25 and 0.33 (BOLIGON et al., 2009; NOBRE et al., 2003; ALBUQUERQUE & MEYER, 2001). Despite the low frequency of dystocia in Nellore cows, BW has been recorded and used to monitor genetic trend. One selection strategy of the breeding programs in Brazil has been to preferentially use sires with higher EBVs for weaning and yearling weights, but with low or close to average EBVs for BW (ALIANÇA, 2011). The identification of major genes and variants affecting

multiple weight and carcass traits or influencing BW alone would be of help to balance these conflicting goals, because BW is positively correlated with weaning and yearling weights (BOLIGON, 2009).

In the past two decades, linkage studies attempting to map quantitative trait loci (QTLs) affecting weight, growth, or stature in cattle have been published (e.g. SPELMAN et al., 1999; KNEELAND et al., 2007; MALTECCA et al., 2009; MCCLURE et al., 2010; COLE et al., 2011). The release of the reference bovine genome (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al., 2009), the discovery of common single nucleotide polymorphisms (SNPs) across breeds (THE BOVINE HAPMAP CONSORTIUM et al., 2009; MATUKUMALLI et al., 2009), and the availability of high-throughput microarrays have enhanced the process of mapping loci that affect complex traits. This has led to several population-based investigations of associations between weight/growth/height phenotypes with genome-wide variants in different cattle breeds (MCCLURE et al., 2010; SNEILLING et al., 2010; PAUSCH et al., 2011; PRYCE et al., 2011; NISHIMURA et al., 2012).

In particular, two regions of the bovine genome associated with stature and growth have been highlighted recently. The first, located on *Bos taurus* (BTA) autosome 6 (SNEILLING et al., 2010; NISHIMURA et al., 2012), shelters the orthologous genes *NCAPG* and *LCORL*, which have been also found to be associated with adult height in humans (GUDBJARTSSON et al., 2008; WEEDON et al., 2008). The second, located on BTA14 (MCCLURE et al., 2010; PAUSCH et al., 2011; PRYCE et al., 2011; NISHIMURA et al., 2012), contains the genes *PLAG1*, *CHCHD7*, *RDHE2*, *MOS*, *RPS20*, *LYN*, *PENK* and *TGS1*, that were previously found to affect stature in both cattle and humans (PRYCE et al., 2011; GUDBJARTSSON et al., 2008; LETTRE et al., 2008; KARIM et al., 2011; LITTLEJOHN et al., 2012). Importantly, the majority of the genome-wide association studies (GWAS) reported in literature were conducted in the humpless subspecies of cattle (*Bos taurus*, known as taurine cattle), and GWAS in the humped bovine subspecies (*Bos indicus*, often referred as indicine or zebu cattle) are only now emerging, especially because the first SNP microarrays were optimized for taurine cattle (MATUKUMALLI et al., 2009).

In this paper, results from a genome-wide scan for SNPs associated with BW variation in Nellore cattle using EBVs of progeny-tested Brazilian bulls are reported.

As EBVs take into account information from performance of the individual, progeny and parents, pedigree relationships, and systematic management and environmental factors, they can be used as composite phenotypes for proceeding with association analyses (see, e.g., GARRICK et al., 2009). The objective of this study was to identify putative SNP associated with differences in BW and to explore the genomic regions around them to unravel prospective functional relationships among weight, fertility, and growth/size traits.

3. Material and methods

3.1. Estimated breeding values

Estimated breeding values for BW were obtained from routine genetic evaluations using performance and pedigree data from the Aliança database (ALIANÇA, 2011), containing data from different commercial Nellore breeding programs, including more than 250 farms distributed across Brazil and Paraguay. The genetic evaluation for BW was calculated using a subset of that data that included 542,918 animals, born from 1985 to 2011, and distributed in approximately 5,000 distinct contemporary groups. These data were collected in 243 grazing-based herds in Brazil. Estimated breeding values were obtained using an animal model that included fixed effects for the age of dam at calving and contemporary group (defined as animals from the same herd, born in the same year and season, and belonging to the same management group at birth, and sex), as well as random effects that include direct additive genetic, maternal additive genetic, maternal permanent environmental and residual error effects. The variance ratios required to solve the mixed model equations were computed based on restricted maximum likelihood (REML) estimates of the variance components from previous studies in this population. Only EBVs of progeny-tested bulls whose accuracy (i.e., square root of reliability, calculated based on prediction error variance estimates) was ≥ 0.50 were used for sample collection and genotyping (described later). The majority of the bulls were used under artificial insemination service.

3.2. Genotyping, informativeness, and quality assurance

A total of 654 progeny-tested Nellore bulls were genotyped with the Illumina® BovineHD Genotyping BeadChip assay, according to the manufacturer's protocol. Genotype calls (i.e. successfully determined genotypes) were defined as genotypes with GenCall Scores greater than 0.70, using the validated standard cluster file provided by the manufacturer. As chromosomes X, Y and mtDNA present different mode of inheritance from the rest of the genome, only autosomal markers with unique genomic coordinates were included into the analyses. After this initial screening, potential duplicated samples were determined by calculating the proportion of alleles identical by state (IBS) shared between all pairs of individuals. Any pair of samples with IBS ≥ 0.95 for 2,000 randomly sampled markers was considered unexpected duplicates, and resulted in the exclusion of both members of the pair. Individual SNPs were removed from the dataset if they did not exhibit: 1) minor allele frequency (MAF) greater than or equal to 0.02, 2) Fisher's exact test *P*-value for Hardy-Weinberg Equilibrium (HWE) greater than or equal to 1×10^{-5} (i.e. extremely deviating from HWE, suggesting potential genotyping error) or 3) Call rate (CR_{SNP}) of at least 98%. After the SNP pruning, individuals exhibiting call rate (CR_{IND}) below 90% were also removed. These procedures and many others described later were performed in the *R v2.15.0* environment (R DEVELOPMENT CORE TEAM, 2008), using combinations of functions from the *R base*, locally developed scripts, and the *GenABEL v1.7-2* package (AULCHENKO et al., 2007b).

3.3. Assessment of population substructure

Sires genotyped in this study were known to belong to one of two major breeding program subgroups in the Aliança database (ALIANÇA, 2011) that have different selection objectives. One group emphasizes selection for weaning and yearling weight (subgroup 1) and the other emphasizes selection for fertility and carcass traits (subgroup 2). Thus, genetic stratification was expected and therefore population substructure was evaluated by performing a Principal Coordinates

Analysis (PCoA). Pair-wise genomic kinship coefficients for all subjects under study were calculated first, following Amin et al. (2007) and Astle and Balding (2009):

$$\hat{f}_{i,j} = \frac{1}{L}\sum_{l=1}^{L} \frac{(g_{l,i}-p_l)(g_{l,j}-p_l)}{p_l(1-p_l)} \quad \textbf{[1]}$$

where $\hat{f}_{i,j}$ is the estimated genomic kinship between individuals i and j, L is the total number of loci used for the calculation, p_l is the reference allele frequency for locus l, and $g_{l,i}$ and $g_{l,j}$ are the locus l genotypes for individuals i and j, respectively (coded as 0, 1 or 2 reference alleles). Calculations were based on 10,000 randomly sampled markers using *GenABEL* (AULCHENKO et al., 2007b). The calculated genomic kinship coefficients within the yielded n x n symmetric matrix (where n is the total number of samples) were then transformed to squared Euclidean distances, and the dissimilarities between the subjects within the matrix were captured in $n-1$ dimensional spaces of n observations (eigenvectors), via classical multidimensional scaling (MARDIS, 1978).

A clustering analysis was applied to the two eigenvectors that explained the largest proportion of the data variance using the k-means algorithm (HARTIGAN & WONG, 1979) implemented in *R* (R DEVELOPMENT CORE TEAM, 2008). Individuals were clustered into 2 groups, and the association between the prior information on breeding program and the k-means clustering results was tested using Pearson's χ^2 with Yates' continuity correction in order to see if the algorithm could reproduce the known breeding programs subgroups. Additionally, an F-test for homogeneity of variance between subpopulations and a t-test for difference between subpopulation means, defining subpopulations either as k-means assignments or breeding program of origin, was performed to determine if there was confounding due to stratification.

3.4. Association analysis

In order to reduce computation time, the ideas of Aulchenko et al. (2007a) were abstracted and a three-step association analysis was performed. In the first step, a linear regression using the weighted least squares method with weights equal to the squared accuracy (i.e., reliability) of the EBVs was applied. By weighting the EBVs by their respective accuracies, the uncertainty around the estimates was taken into account when estimating the regression parameters. The following model was fitted:

$$y_i = \mu + \sum_{j=1}^{n} \beta_j X_{ij} + \varepsilon_i \textbf{ [2]}$$

Where y_i is the EBV of sire i, μ is the overall mean, X_{ij} is value i (corresponding to sire i) in the eigenvector j calculated in the PCoA, β_j is the estimated effect of eigenvector j, and ε_i is the residual effect for animal i. Only eigenvectors significantly ($P < 0.05$) correlated with the dependent variable, as assessed by Pearson correlations, were included in the model. Next, the residuals were obtained from the fitted model in **[2]**:

$$y_i^* = y_i - \hat{y}_i \textbf{ [3]}$$

and had their homoskedasticity and normality tested by using the studentized Breusch-Pagan test and the Shapiro-Wilk test, respectively. Then, these residuals were used as the new dependent variable for a single-marker linear regression:

$$y_i^* = \mu + \beta_g g_i + \varepsilon_i \textbf{ [4]}$$

where β_g is the marker regression coefficient (i.e., the allele substitution effect of the SNP) and g_i is the genotype (0, 1 or 2) of the sire i. For each SNP, β_g and its respective standard error (SE_g) were estimated using ordinary least squares. The

association between the SNP and the trait was assessed via a test statistic, calculated as:

$$T^2 = \frac{\widehat{\beta}_g^2}{SE_g^2} \textbf{ [5]}$$

The test statistics are assumed to asymptotically follow a χ^2 distribution with one degree of freedom under the null hypothesis. To assess the validity of this assumption, the deviation of the distribution of the test statistics from the expected theoretical quantiles was examined via 1) a quantile-quantile (Q-Q) plot, and 2) calculation of the inflation/deflation factor:

$$\lambda = \frac{median(T^2)}{0.456} \textbf{ [6]}$$

If $\lambda < 1.1$, the inflation was considered acceptable, and the Genomic Control (GC) correction was applied to adjust for that inflation (DEVLIN & ROEDER, 1999). Then, *P*-values were derived from the χ^2 cumulative distribution function for the corrected test statistics. Finally, markers within the smallest 0.1% *P*-value percentile (i.e., most significant) were considered for re-analysis with the full model:

$$y_i = \mu + \sum_{j=1}^{n} \beta_j X_{ij} + \beta_g g_i + \varepsilon_i \textbf{ [7]}$$

The EBVs were again weighted by their respective accuracies. The conservative Bonferroni adjustment for multiple testing (α = 0.05 / N, where N is the number of tests, i.e., number of SNPs) was used to reject the null hypothesis ($\beta_g = 0$, i.e. there is no association between the SNP and the EBVs), which resulted in an adjusted significance of $\alpha = 1.15 \times 10^{-7}$.

3.5. Exploratory view of significant SNPs

For any peak crossing the Bonferroni significance threshold, the estimated regression parameters were reported. For the most significant SNP, a 95% confidence interval (CI) for the estimated allele substitution effect size ($\hat{\beta}_g$) was calculated, and the percentage of the EBV variance explained was calculated as:

$$\%\hat{\pi}_{\sigma^2} = \frac{2pq\hat{\beta}_g^2}{S^2} \text{x } 100 \quad \textbf{[8]}$$

Where p and q are the allele frequencies and S^2 is the sample EBVs variance. The upper and lower limits of the estimated 95% CI for β_g were used to derive a 95% CI for $\%\hat{\pi}_{\sigma^2}$.

The genomic region containing the most significant SNP of a peak was explored by inspecting a 1 Mb window around the location of this SNP using the *BioMart tool* and the *Ensembl genes 69* database (KINSELLA et al., 2011) to interrogate 500 kb to each side of the marker using the *UMD v3.1* assembly. The cattle *QTLdb* database (HU et al., 2013) was also examined to find out if the significant SNP mapped against any previously described bovine QTL. Additionally, the alignments of the *UMD v3.1* assembly sequence of the 1 Mb window against the human (*Homo sapiens*, *GRCh37* assembly), pig (*Sus scrofa*, *Sscrofa10.2* assembly) and mouse (*Mus musculus*, *GRCm38* assembly) genome builds were inspected in the *Ensembl Comparative genomics alignments* and *Comparative genomics synteny* tools in order to determine if any homologous genes were present in the putative region.

4. Results

4.1. Genotype informativeness and quality control

From the initial set of 777,961 SNPs, 42,669 (5.5%) were non-autosomal markers. Fifty four autosomal SNPs with redundant genomic coordinates were identified and excluded from further analyses. The IBS check revealed no unexpected sample duplicates. A total of 223,309 (30.4%) markers were excluded due to MAF < 0.02. The number of SNPs excluded due to CR_{SNP} < 0.98 and Fisher's exact test *P*-value for HWE < $1x10^{-5}$ were 122,611 (16.7%) and 13,194 (1.8%), respectively. Five individuals were removed due to low CR_{IND}. The final dataset included data for 649 individuals and 434,020 SNPs.

4.2. Descriptive statistics of dependent variables

Based on the results for the Shapiro-Wilk test, there was no evidence that EBVs deviated from normality (*P* = 0.415), and no outliers were observed. Average accuracy was 0.87 ± 0.11, with minimum, median and maximum accuracies of 0.51, 0.91 and 0.99, respectively. After fitting the regression in formula **[2]**, there was no evidence against the hypotheses of normally distributed (*P* = 0.57) and homoskedastic residuals (studentized Breusch-Pagan test, *P* = 0.11). These findings suggest that the dependent variables used were reliable and did not violate possible assumptions of the statistical analyses used hereafter.

4.3. Population substructure

The PCoA revealed genetic stratification among the Nellore samples (Figure 1). After using k-means clustering to assign individuals to two different groups according to their coordinates in the PCoA, a highly significant association (*P* = $5.41x10^{-34}$) between the k-means assignments ($n_{cluster1}$= 531, $n_{cluster1}$ = 118) and breeding program subgroups ($n_{subgroup1}$ = 352, $n_{subgroup2}$ = 297) was found. For the k-means groups, BW average was 0.434 ± 1.306 in cluster 1 and -0.216 ± 1.295 in

cluster 2. For the breeding program subgroups, the trait averages in subgroup 1 and subgroup 2 were 0.683 ± 1.276 and -0.119 ± 1.255, respectively. When considering either k-means clusters or breeding program subgroups as subpopulation labels, the EBVs showed homogeneity of variance ($P > 0.05$), but the trait mean was significantly different between groups ($P = 1.21 \times 10^{-6}$ and $P = 4.37 \times 10^{-15}$ for k-means and breeding program, respectively). Thus, population substructure was a potential confounder in the genotype-EBV association analysis, which justified the inclusion of eigenvectors from the PCoA as fixed effects in the linear model.

4.4. Association analysis

A total of 32 eigenvectors from the PCoA were significantly correlated with the phenotypes, which together explained 15.23% of the genotypic variability. Residuals from the weighted regression on these significant eigenvectors were used as the dependent variable for the SNP association analysis. The Q-Q plot (Figure 2) showed that the deviation of the observed test statistics from the theoretical quantiles was mild and acceptable ($\lambda = 1.002831$), and the values were adjusted for the inflation factor via GC. The T^2 values deviating from the expected values were interpreted as SNPs departing from the null hypothesis. The genome-wide deflated *P*-values are shown in Figure 3.

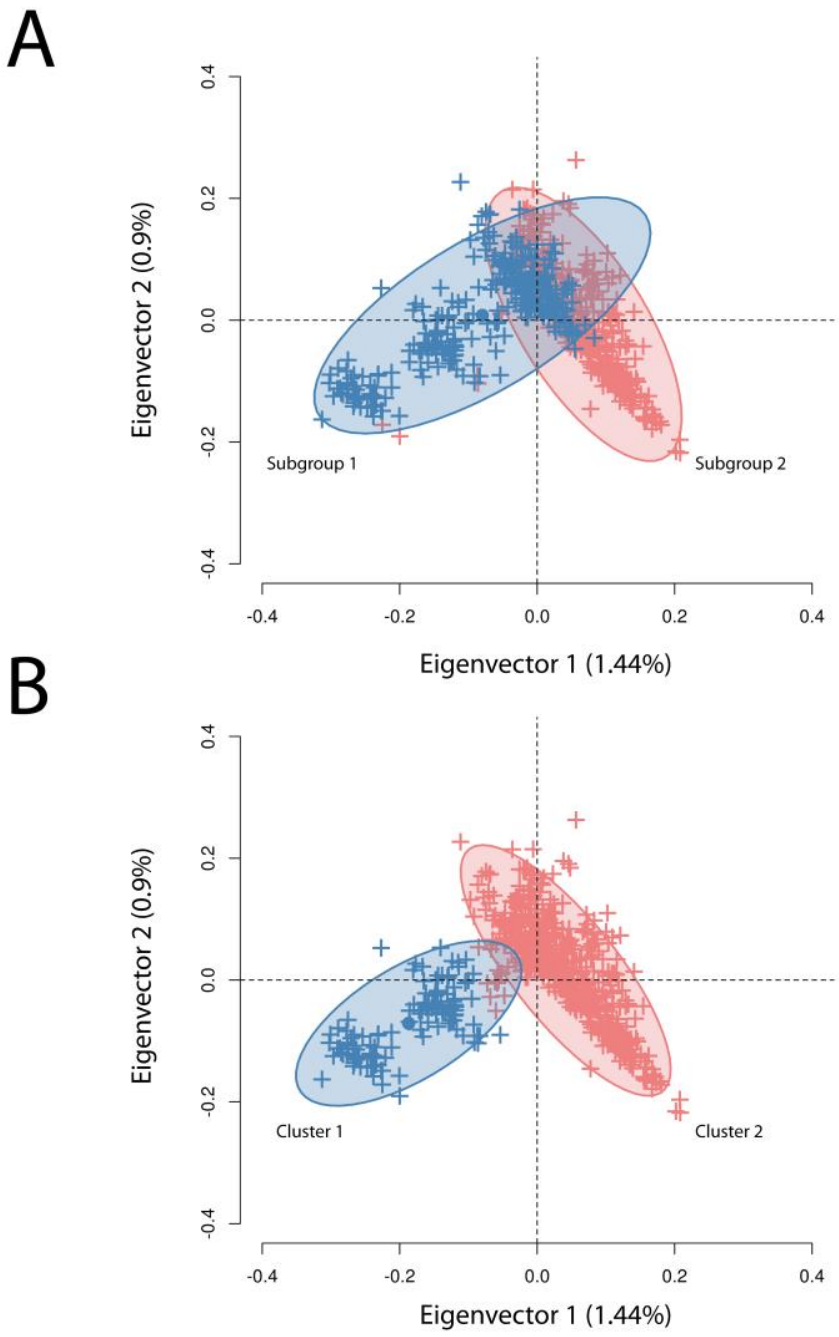

Figure 1. Principal Coordinates Analysis based on the genomic kinship coefficient. Percentages inside brackets correspond to the variance explained by each respective eigenvector. Each '+' represents an individual and ovals are 95% inertia ellipses. A) Subjects colored according to breeding program subgroups. B) Subjects colored according to k-means clustering results.

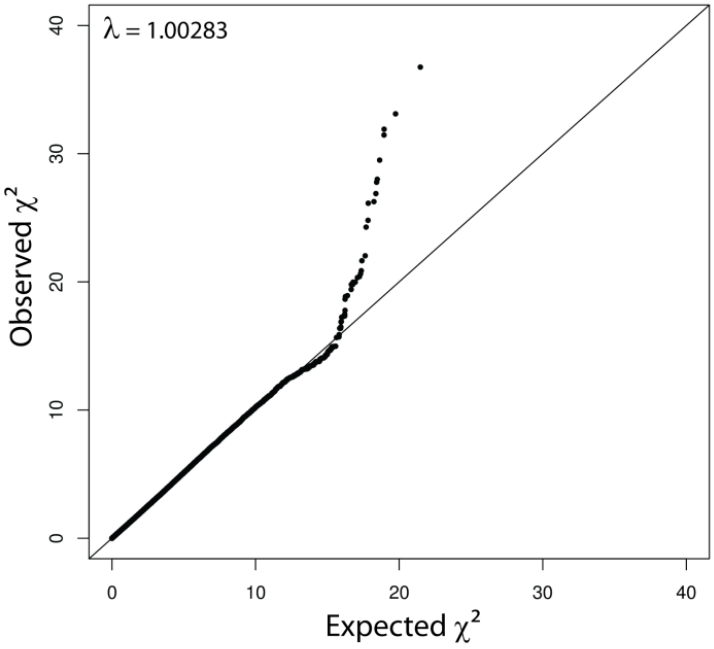

Figure 2. Quantile-quantile plot for the test statistics (χ^2) used in the association analysis.

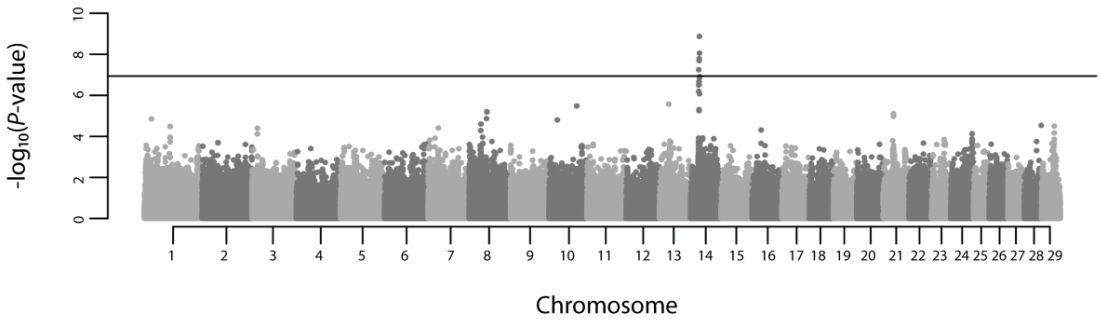

Figure 3. Manhattan plot of genome-wide -$\log_{10}$(*P*-values) for birth weight estimated breeding values in Nellore cattle. The horizontal line represents the Bonferroni significance threshold (α = 1.15x10^{-7}).

A peak crossing the boundary for Bonferroni significance (α = 1.15x10^{-7}) was detected on BTA14, comprising 5 SNPs (Table 1) which were highly linked (mean r^2 = 0.728 ± 0.12). The most significant SNP (rs133012258, P_{GC} = 1.34x10^{-9}), located at BTA14:25376827, had an estimated allele substitution effect of 0.452 kg (i.e., for each extra A allele, the BW breeding value is expected to increase 0.452 kg), with lower and upper limits for the 95% CI of 0.306 kg and 0.598 kg, respectively, and the percentage of the variance in sires EBVs explained by the SNP was 4.62% (with a 95% CI of 2.12-8.09%). The overall rs133012258 A allele frequency was 0.274, whereas the breeding program subgroups 1 and 2 had frequencies of 0.351 and 0.184, respectively. Figure 4 shows the distribution of the EBVs (in standard deviations) for the three genotype classes of rs133012258. In both Illumina TOP and Forward allele notation, the AB correspondence for rs133012258 was A = A and B = G.

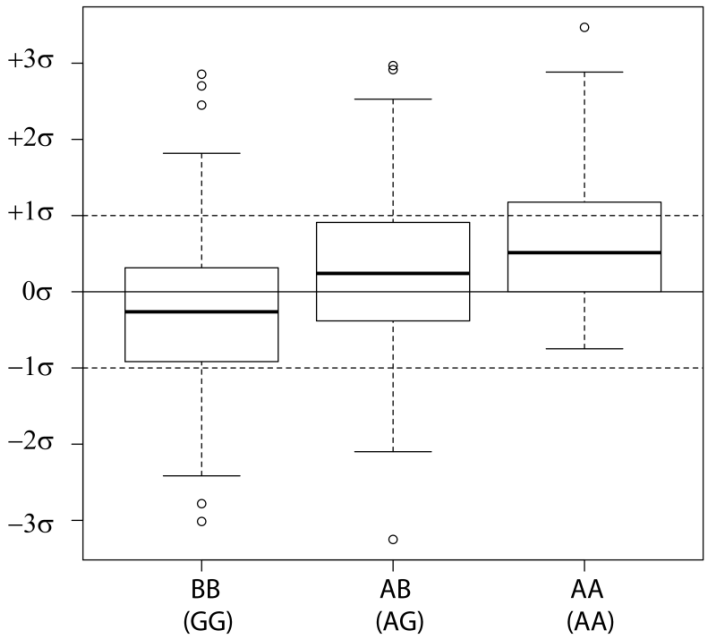

Figure 4. Box plots for the birth weight estimated breeding values according to rs133012258 genotypes. Values in the y axis are expressed in terms of standard units.

Table 1. Summary of parameters and statistics estimated for the identified significant SNPs

Ensembl variant ID	Illumina probe ID	BTA14 Position (bp)	n	Effect allele	Allele frequency	CR [a] (%)	HWE *P*-value	β[b](kg)	SE	T^2 [c]	$GC-T^2$ [d]	*P*-value
rs133012258	BovineHD1400007343	25376827	649	A	0.274	100.00	0.920	0.452	0.074	36.858	36.753	$1.34x10^{-9}$
rs41627948	BovineHD1400007374	25504073	648	B	0.181	99.85	0.110	0.522	0.090	33.202	33.108	$8.72x10^{-9}$
rs42646720	BovineHD1400007144	24590812	649	B	0.243	100.00	0.390	0.457	0.081	31.999	31.909	$1.62x10^{-8}$
rs136764901	BovineHD1400007159	24651537	641	B	0.244	98.77	0.200	0.459	0.082	31.546	31.457	$2.04x10^{-8}$
rs136287861	BovineHD1400006765	23313228	645	A	0.272	99.38	0.766	0.408	0.075	29.580	29.497	$5.60x10^{-8}$

[a] CR = Call rate
[b] β = Estimated allele substitution effect
[c] T^2= Chi-squared statistics for β: $\chi 2 = T^2 = \frac{\beta^2}{SE^2}$
[d] GC-T^2= Corrected Chi-squared statistics for β. Genomic control correction was performed by dividing the $\chi 2$ statistics by the distribution inflation/deflation factor estimate (λ = 1.002831).

Table 2. List of genes within the 1 Mb region surrounding the most significant SNP (rs133012258)

Gene	Ensembl ID	BTA14 coordinates	Distance from SNP (kb)	Strand	HSA8 homology	SSA4 homology	MMA4 homology	Description
U6	ENSBTAG00000043923	25492090:25492184	115.4	+	No homologues	No homologues	No homologues	U6 spliceosomal RNA
PENK	ENSBTAG00000004924	25218586:25222991	153.8	-	ENSG00000181195	ENSSSCG00000006243	ENSMUSG00000045573	Proenkephalin-A
IMPAD1	ENSBTAG00000015637	25544907:25560879	168.1	-	ENSG00000104331	ENSSSCG00000006242	ENSMUSG00000066324	Inositolmonophosphatase 3
SDR16C6	ENSBTAG00000040321	25153583:25179651	197.2	-	No homologues	No homologues	ENSMUSG00000071019	Short-chain dehydrogenase / reductase family 16C
SDR16C5 (RDHE2)	ENSBTAG00000018570	25105062:25117554	259.3	-	ENSG00000170786	ENSSSCG00000006245	ENSMUSG00000028236	Epidermal retinol dehydrogenase 2
Unknown	ENSBTAG00000039031	25067486:25067823	309.0	-	No homologues	No homologues	No homologues	Uncharacterized protein
CHCHD7	ENSBTAG00000033284	25052885:25058779	318.0	+	ENSG00000170791	ENSSSCG00000006246	ENSMUSG00000042198	Coiled-coil-helix-coiled-coil-helix domain-containing protein 7
PLAG1	ENSBTAG00000004022	25007291:25009296	367.5	-	ENSG00000181690	ENSSSCG00000006247	ENSMUSG00000003282	Pleiomorphic adenoma gene 1
MOS	ENSBTAG00000019145	24975950:24976948	399.9	-	ENSG00000172680	ENSSSCG00000006248	ENSMUSG00000078365	V-mos Moloney murine sarcoma viral oncogene homolog
U1	ENSBTAG00000028889	24970516:24970679	406.1	+	No homologues	No homologues	No homologues	U1 spliceosomal RNA
RPS20	ENSBTAG00000019147	24955079:24956324	420.5	-	ENSG00000008988	ENSSSCG00000006249	ENSMUSG00000028234	40S ribosomal protein S20
snoU54	ENSBTAG00000045097	24955769:24955835	421.0	-	No homologues	No homologues	No homologues	Small nucleolar RNA U54
LYN	ENSBTAG00000020034	24847257:24920713	456.1	+	ENSG00000254087	ENSSSCG00000006250	ENSMUSG00000042228	Tyrosine-proteinkinase Lyn

One of the 5 significant SNPs, rs42646720 (BTA14:24590812, P_{GC} = 1.62x10^{-8}), is located within intron 2 of the gene Kell blood group complex subunit-related family, member 4 (*XKR4* or *KIAA1889*, ENSBTAG00000044050). Thirteen genes were found within 500 kb of the most significant SNP (Figure 5), including the human height-associated orthologous genes *PLAG1*, *CHCHD7*, *MOS*, *RPS20*, *LYN*, *RDHE2* (*SDR16C5*) and *PENK* (Table 2). The bovine reference genome sequence of this region was found to have high identity with human (*Homo sapiens* - HSA), pig (*Sus scrofa* - SSC) and mouse (*Mus musculus* - MMU) autosomes 8 (HSA8), 4 (SSC4) and 4 (MMU4), respectively (Figure 6), and the majority of the genes had homology across species (Table 2). The most significant SNP also overlapped 28 QTLs previously reported in the literature by linkage mapping studies using different cattle breeds (Table 3), including QTLs for birth weight, mature height, carcass weight, stature, pre-weaning average daily gain, calving ease and gestation length.

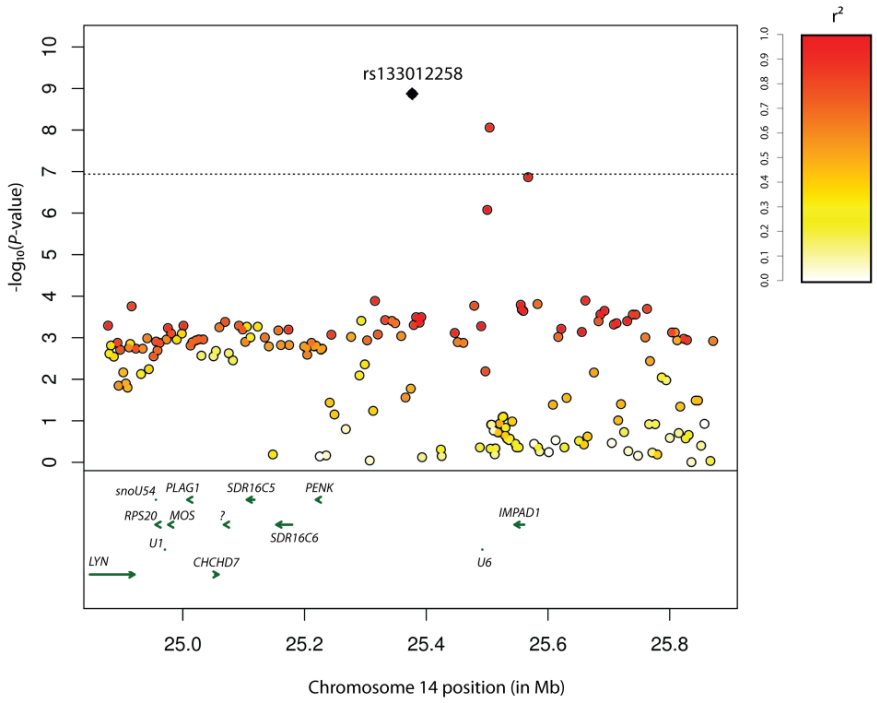

Figure 5. Regional association plot for birth weight in the 1 Mb window around rs133012258. Upper box: each dot represents a SNP, and its color heat the degree of linkage disequilibrium with rs133012258 (black diamond). The horizontal dashed line represents the Bonferroni significance threshold (α = 1.15x10^{-7}). Lower box: genes (green arrows; right-handed = positive

strand, left-handed = negative strand) within the region in the *UMD v3.1* assembly.

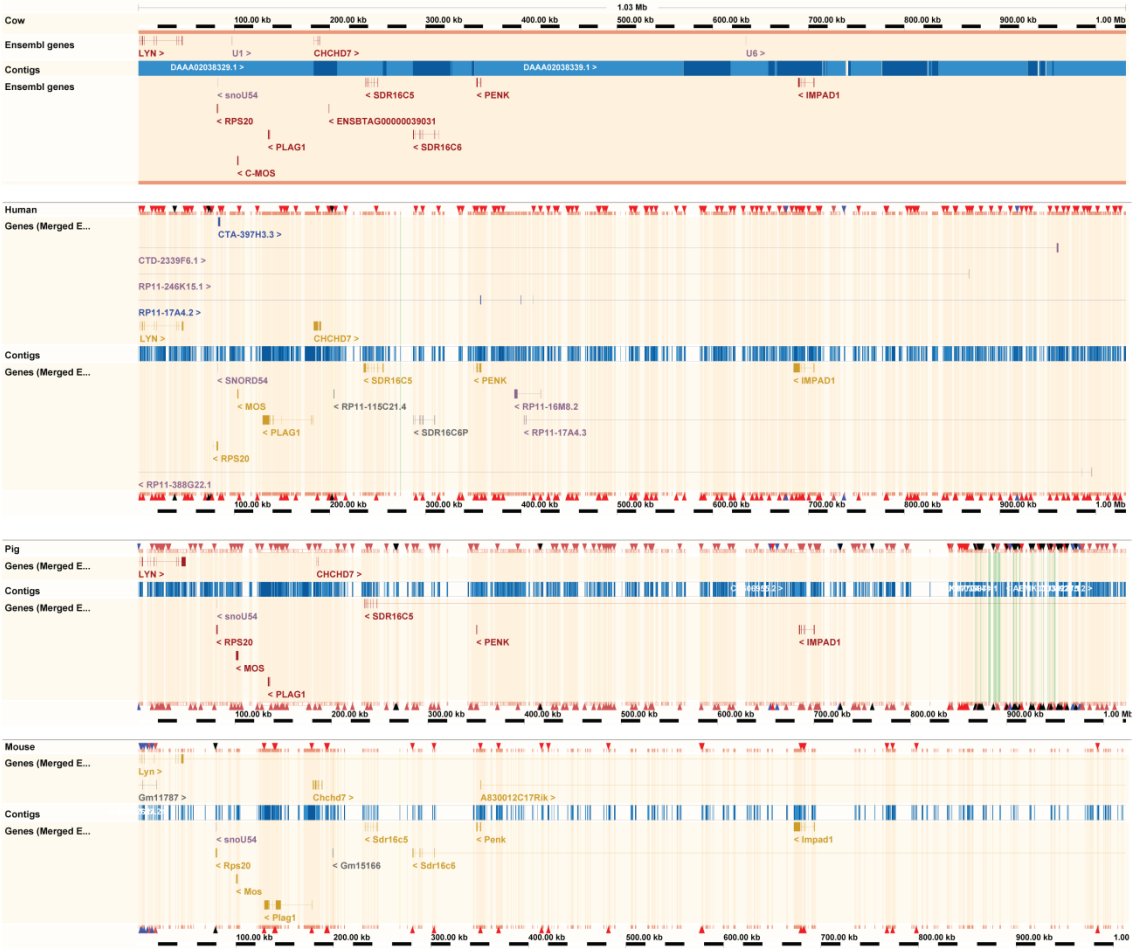

Figure 6. Ensembl alignments of *UMD v3.1* sequence for the 1 Mb region surrounding rs133012258. The bovine reference genome sequence was aligned against (from top to bottom) the human (*GRCh37* assembly), pig (*Sscrofa10.2* assembly) and mouse (*GRCm38* assembly) genome builds. Gene colors: yellow - merged Ensembl/Havana, red - protein coding, blue - processed transcript, grey - pseudogene, purple - RNA gene. Triangles: black - breakpoint between different chromosomes, blue - inversion in chromosome, brown - breakpoint on chromosome, red - gap between two underlying slices.

Table 3. QTLdb hits within the 1 Mb region surrounding the most significant SNP (rs133012258)

Trait	BTA14 coordinates	QTLdb ID	PubMed ID
Body weight (birth)	6311565:71762521	5375	19016677
Height (mature)	19204282:42398519	10962	20477797
Carcass weight	25224396:30870876	1375	16151698
	10808022:28658498	10960	20477797
Stature	25219037:65017465	4613	10575619
Pre-weaning average daily gain	25224396:35530275	2630	15537758
Calving ease (maternal)	19204282:28658498	10959	20477797
Gestation length	6311565:30372479	5374	19016677
	17512260:48646289	5385	19016677
Rump angle	25224396:29928419	1592	16230715
Longissimus muscle area	19204282:42398519	10964	20477797
Fat thickness at the 12th rib	19204282:28658498	10961	20477797
Marbling score	16670076:73064076	1334	14677852
Abnormal flavor intensity	5565085:28658498	4833	18254735
Tick resistance	5545944:77366190	9917	17894560
Milk yield	25372161:25525285	6209	18650300
	16243029:28716621	3608	12729552
Milk fat yield (or percentage)	13401044:35992517	2733	9691050
	1641277:25448723	3408	12605852
	25224396:29928419	2676	14762090
	13401044:35992517	2732	9691050
Milk protein yield (or percentage)	1641277:56300551	3413	12605852
	9479897:44651695	2604	12778594
	1641277:81189386	10099/10100/10101	18298934
Somatic cell score	13401044:35992517	2734	9691050
	13401044:35992517	2776	14556700
	20567087:44651695	4884	17954769
Clinical mastitis	9479131:44651695	3177	14762087

5. Discussion

Five SNPs on BTA14 were identified as associated with BW in Nellore cattle ($P < 1.15 \times 10^{-7}$), whose surrounding region has been shown to contain many QTLs, genes and variants affecting stature-related traits in cattle by several independent studies (SPELMAN et al., 1999; KNEELAND et al., 2004; MALTECCA et al., 2009; MCCLURE et al., 2010; PAUSCH et al., 2011; PRYCE et al., 2011; NISHIMURA et al., 2012). More particularly, the genes *PLAG1*, *CHCHD7*, *RDHE2*, *MOS*, *RPS20*, *LYN* and *PENK* have been found to influence both human and cattle height (PRYCE

et al., 2011; NISHMURA et al., 2012; GUDBJARTSSON et al., 2008; WEEDON et al., 2008; LETTRE et al., 2008; KARIM et al., 2011; LITTLEJOHN et al., 2012).

The BTA14 region pointed out by the present study has also been shown to be associated with reproductive traits. Cole et al. (2011) reported a QTL on BTA14 associated with stillbirth, which also has been associated with body size in dairy cattle (JOHANSON & BERGER, 2003; COLE et al., 2009), but found no effect on stature or other conformation traits on that chromosome. The region also associates with many fertility and growth-related traits in the indicine breed Brahman, for example scrotal circumference (FORTES et al., 2012a; FORTES et al., 2012b), age at the first corpus luteum (FORTES et al., 2012a; HAWKEN et al., 2012), blood levels of insulin-like growth factor 1 (IGF1) (FORTES et al., 2012b; HAWKEN et al., 2012) and hip height (HAWKEN et al., 2012).

A significant SNP was found within intron 2 of the *XKR4* gene in the present study. Lindholm-Perry et al. (LINDHOLM-PERRY et al., 2012) identified five SNPs near *XKR4* associated with feed intake and gain in crossbred steers. Bolormaa et al. (2011) found five SNPs in a narrow region of BTA14 encompassing *XKR4* associated with rump fat thickness measured at the P8 position (CHILLP8) in seven breeds of cattle, including taurine, indicine and composite breeds. The authors found that four of these SNPs were also associated with CHILLP8 in a confirmatory sample of 1,338 animals, including Angus, Hereford and Brahman cattle. Furthermore, Porto-Neto et al. (2012) performed a replication study using samples of Belmont Red, Santa Gertrudis and Brahman animals genotyped for SNPs within XKR4 and found that although the SNP effect may vary depending on the breed, the variant rs42646708 (BTA14:24573257) explain around 1.3% of CHILLP8 variance in cattle. This SNP is also located within intron 2 of *XKR4*, only 17.6 kb apart from the intronic SNP detected in the present study, which strongly suggests *XKR4* as a candidate gene for being further explored in future studies of weight and carcass traits in Nellore cattle.

The most significant SNP (rs133012258, $P_{GC} = 1.34 \times 10^{-9}$) was found to explain 4.62% of the variance in sires EBVs, with a 95% CI of 2.12-8.09%. One hundred and eighty loci associated with human adult height explain only 10% of the phenotypic variance together, while individual loci account for 0.4% or less (LANGO et al., 2010). SNPs analyzed by (PRYCE et al., 2011) within a nearby BTA14 region

explain from 0.29 to 2.53% of the bovine stature variability, and the quantitative trait nucleotides (QTN) spanning *MOS*, *CHCHD7* and *PLAG1* described by (KARIM et al., 2011) explain from 1.10 to 3.50% of height in Jersey and Holstein breeds. Furthermore, the genome-wide survey performed by (PAUSCH et al., 2011) provided strong evidence for two QTL on BTA14 and BTA21 that together explain at least 10% of the variation of EBVs for calving ease in the German Fleckvieh.

Considering that multiple stature-related traits are governed by variants with small effects, and that the genomic region identified in this study has been previously found to be associated with several of these traits, the putative SNP detected in the present analysis can be considered as a marker in linkage disequilibrium (LD) with major untyped (i.e., not probed by the SNP assay used) causative variants affecting BW and other height-associated traits in Nellore cattle, and further studies would be needed to determine if the QTNs reported by Karim et al. (2011) are also segregating in the Nellore population. Also, future investigations are needed to better characterize the effect of nearby SNPs on other weight and carcass traits in Nellore cattle, as it is not clear yet how the putative pleiotropic effect of these variants would be used towards balancing conflicting selection goals for birth, weaning and yearling weights. Although we cannot confirm that the allele substitution effects of these SNPs work in the same direction for all three traits, because only birth weight was analyzed here, these findings suggest that the SNPs identified would be key polymorphisms to be monitored over time. In a scenario where the SNP effects have the same direction in all three traits, one strategy could be avoiding strong positive selection or drifting of the allele that contributes to higher BW EBVs, and identify and promote positive selection of other variants that have effects on weaning and yearling weights only.

The high identity found in the alignment of this BTA14 region against other mammalian species genomes suggests that these orthologous genes are located in a conserved syntenic block which may have arisen and been maintained after speciation from a common ancestor of the mammal clade. Moreover, the evidence for variants associated with growth and stature within this BTA14 region in both taurine and zebu cattle raises two hypotheses: 1) these variants have been introgressed into Nellore via historical admixture with taurine Creole cattle in the maternal line, and was maintained in the breed in spite of several generations of

backcrossing; 2) these are ancient polymorphisms, probably already segregating in the founder population of wild Aurochs (*Bos primigenius*) before subspecies formation.

Regarding functional meaning, the set of genes reported participate in diverse growth and tumor development mechanisms. Among these genes, *PLAG1* is the most appealing functional candidate. It is an oncogene that encodes a transcription factor broadly expressed during fetal development, but is down-regulated at birth (KARIM et al., 2011). It interacts with several growth factors controlling body size, including IGF2 (VAN DYCK et al., 2007). In addition, *PLAG1* knock-out mice have been shown to have marked growth retardation and reduced fertility (HENSEN et al., 2004). In a replication study, Littlejohn et al. (2012) confirmed the findings reported by Karim et al. (2011), demonstrating association of growth rate and early life and peripubertal body weight with *PLAG1* polymorphisms, supporting its status as a key regulator of mammalian growth.

The lack of significant association between BW and SNPs within other previously described weight- and height-related chromosome regions in the present study should not be interpreted as a lack of existence of true association, but rather it might be due to limitations specific to this study. Firstly, because complex trait mapping requires large sample sizes and only 649 bulls were analyzed here. Secondly, the significance level adopted was highly stringent, which may have caused inflation of type II errors. In spite of these limitations, it was possible to demonstrate that a well-characterized chromosome region affecting human and taurine cattle stature also associates with BW in a zebu breed. The release of a *Bos indicus* reference genome assembly, as well as the application of re-sequencing and replication studies would help improve resolution to narrow down the genomic region as close as possible to the true causative variants.

6. Conclusions

This study is believed to be the first genome-wide association study applying a high-density SNP panel to identify putative chromosome regions affecting birth weight in zebu cattle. The findings presented, which are strongly supported by the

literature, point to orthologous genes already known to affect growth- and stature-related traits in both humans and cattle, which may shelter ancient polymorphisms responsible for variation in those traits since before cattle subspecies divergence.

7. Acknowledgements

We thank Guilherme Penteado Coelho Filho and Daniel Biluca for technical assistance in sample acquisition. This research was supported by: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – process 560922/2010-8 and 483590/2010-0; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - process 2011/16643-2 and 2010/52030-2; Next-Generation BioGreen 21 Program (No. PJ008196), Rural Development Administration, Republic of Korea. Mention of trade name proprietary product or specified equipment in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the authors or their respective institutions.

8. Author’s contributions

J. F. Garcia conceived and led the coordination of the study. J. Sölkner, J. McEwan, J. B. Cole, C. P. Van Tassell, F. S. Schecnkel, M. V. G. B. Silva, L. R. Porto-Neto and T. S. Sonstegard contributed to the study design and coordination. A. S. Carmo directed the genotyping work and contributed to the data analysis. R. Carvalheiro and H. H. R. Neves provided EBVs and assisted the data analysis. Y. T. Utsunomiya led the data analysis and the manuscript preparation. M. C. Matos, L. B. Zavarez and A. M. Pérez O’Brien contributed in the data preparation and analysis. All authors read and approved the final manuscript.

9. References

Albuqueque, L. G.; Meyer, K. Estimates of direct and maternal genetic effects for weights from birth to 600 days of age in Nelore cattle. **Journal of Animal Breeding and Genetics**, v. 118, p. 83-92, 2001.

Aliança. Sumário de touros 2011 Aliança Nelore. **GenSys Consultores Associados**, 2011. Available at: <http://www.gensys.com.br/home/show_page.php?id=701>, Accessed on 19 dec. 2012.

Amin, N.; Van Duijn, C. M.; Aulchenko, Y.S. A genomic background based method for association analysis in related individuals. **PLoS One**, 2:e1274, 2007.

Astle, W.; Balding, D. J. Population structure and cryptic relatedness in genetic association studies. **Statistical Science**, v. 24, n. 4, p.451-471, 2009.

Aulchenko, Y. S.; de Koning, D-J.; Haley, C; Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. **Genetics**, v. 177, p. 577-585, 2007a.

Aulchenko, Y. S.; Ripke, S.; Isaacs, A.; Van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. **Bioinformatics**, v. 23, n. 10, p. 1294-1296, 2007b.

Boligon, A. A.; Albuquerque, L. G.; Mercadante, M. E. Z.; Lôbo, R. B. Herdabilidades e correlações entre pesos do nascimento à idade adulta em rebanhos da raça Nelore. **Revista Brasileira de Zootecnia***,* v. 38, n. 12, p.2320-2326, 2009.

Bolormaa, S.; Porto-Neto, L. R.; Zhang, Y. D.; Bunch, R. J.; Harrison, B. E.; Goddard, M. E.; Barendse, W. A genome-wide association study of meat and carcass traits in Australian cattle. **Journal of Animal Science**, v. 89, n. 2, p. 297-309, 2011.

Bourdon, R. M.; Brinks, J. S. Genetic, environmental and phenotypic relationships among gestation length, birth weight, growth traits and age at first calving in beef cattle. **Journal of Animal Science**, v. 55, n. 3, p. 543-553, 1982.

Cole, J. B., VanRaden, P. M.; O'Connell, J. R.; Van Tassell, C. P.; Sonstegard T. S; Schnabel, R. D.; Taylor, J. F.; Wiggans, G. R. Distribution and location of genetic effects for dairy traits. **Journal of Dairy Science**, v. 92, p. 2931-2946, 2009.

Cole, J. B.; Wiggans, G. R.; Ma, L.; Sonstegard, T. S.; Lawlor, T. J.; Crooker, B. A.; Van Tassell, C. P.; Yang, J.; Wang, S.; Matukumalli, L. K.; Da Y. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. **BMC Genomics**, 12:408, 2011.

Cook, B. R.; Tess, M. W.; Kress, D. D. Effects of selection strategies using heifer pelvic area and sire birth weight expected progeny difference on dystocia in first-calf heifers. **Journal of Animal Science**, v. 71, p. 602-607, 1993.

Devlin, B.; Roeder, K. Genomic control for association studies. **Biometrics**, v. 55, p. 997-1004, 1999.

Eriksson, S.; Näsholm, A.; Johansson, K.; Philipsson, J. Genetic parameters for calving difficulty, stillbirth, and birth weight for Hereford and Charolais at first and later parities. **Journal of Animal Science**, v. 82, p. 375-383, 2004.

Fortes, M. R. S.; Lehnert, S. A.; Bolormaa, S.; Reich, C.; Fordyce, G.; Corbet, N. J.; Whan, V.; Hawken, R. J.; Reverter, A. Finding genes for economically important traits: Brahman cattle puberty. **Animal Production Science**, v. 52, p. 143-150, 2012a.

Fortes, M. R. S.; Reverter, A.; Hawken, R. J.; Bolormaa, S.; Lehnert, S. A. Candidate genes associated with testicular development, sperm quality, and hormone levels of inhibin, luteinizing hormone, and insulin-like growth factor 1 in Brahman bulls. **Biology of Reproduction**, 87:58, 2012b.

Garrick, D. J.; Taylor, J. F.; Fernando, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, 41:55, 2009.

Gudbjartsson, D. F.; Walters, G. B.; Thorleifsson, G.; Stefansson, H.; Halldorsson, B. V.; Zusmanovich, P.; Sulem, P.; Thorlacius, S.; Gylfason, A.; Steinberg, S.; Helgadottir, A.; Ingason, A.; Steinthorsdottir, V.; Olafsdottir, E. J.; Olafsdottir, G. H.; Jonsson, T.; Borch-Johnsen, K.; Hansen, T.; Andersen, G.; Jorgensen, T.; Pedersen, O.; Aben, K. K.; Witjes, J. A.; Swinkels, D. W.; den Heijer, M.; Franke, B.; Verbeek, A. L.; Becker, D. M.; Yanek, L. R.; Becker, L. C.; Tryggvadottir, L.; Rafnar, T.; Gulcher, J.; Kiemeney, L. A.; Kong, A.; Thorsteinsdottir, U.; Stefansson, K. Many sequence variants affecting diversity of adult human height. **Nature Genetics**, v. 40, p. 609-615, 2008.

Gutiérrez, J. P.; Goyache, F.; Fernández, I.; Alvarez, I.; Royo, L. J. Genetic relationships among calving ease, calving interval, birth weight, and weaning weight in the Asturiana de los Valles beef cattle breed. **Journal Animal Science**, v. 85, p. 69-75, 2007.

Hartigan, J. A.; Wong, M. A. A K-means clustering algorithm. **Applied Statistics**, v. 28, p. 100-108, 1979.

Hawken, R. J.; Zhang, Y. D.; Fortes, M. R. S.; Collis, E.; Barris, W. C.; Corbet, N. J.; Williams, P. J.; Fordyce, G.; Holroyd, R. G.; Walkley, J. R. W.; Barendse, W.; Johnston, D. J.; Prayaga, K. C.; Tier, B.; Reverter, A.; Lehnert, S. A. Genome-wide association studies of female reproduction in tropically adapted beef cattle. **Journal of Animal Science**, v. 90, p. 1398-1410, 2012.

Hensen, K.; Braem, C.; Declercq, J.; Van Dyck, F.; Dewerchin, M.; Fiette, L.; Denef, C.; Van de Ven, W. J. M. Targeted disruption of the murine PLAG1 proto-oncogene causes growth retardation and reduced fertility. **Development, Growth & Differentiation**, v. 46, p. 459-70, 2004.

Hu, Z. L.; Park, C. A.; Wu, X. L.; Reecy, J. M. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. **Nucleic Acids Research**, D871-9, 2013.

Johanson, J. M.; Berger, P. J. Birth weight as a predictor of calving ease and perinatal mortality in Holstein cattle. **Journal of Dairy Science**, v. 86, n. 11, 3745-3755, 2003.

Karim, L.; Takeda, H.; Lin, L.; Druet, T.; Arias, J. A.; Baurain, D.; Cambisano, N.; Davis, S. R.; Farnir, F.; Grisart, B.; Harris, B. L.; Keehan, M. D.; Littlejohn, M. D.; Spelman, R. J.; Georges, M.; Coppieters, W. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. **Nature Genetics**, v. 43, 405-413, 2011.

Kinsella, R. J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; Kersey, P.; Flicek, P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. **Database (Oxford),** Bar030, 2011.

Kneeland, J.; Li, C.; Basarab, J.; Snelling, W. M.; Benkel, B.; Murdoch, B.; Hansen, C.; Moore, S. S. Identification and fine mapping of quantitative trait loci for growth traits on bovine chromosomes 2, 6, 14, 19, 21, and 23 within one commercial line of Bos taurus. **Journal Animal Science**, v. 82, n. 12, p. 3405-3414, 2004.

Lango Allen, H.; Estrada, K.; Lettre, G.; Berndt, S. I.; Weedon, M. N.; Rivadeneira, F.; Willer, C. J.; Jackson, A. U.; Vedantam, S.; Raychaudhuri, S.; Ferreira, T.; Wood, A. R.; Weyant, R. J.; Segrè, A. V.; Speliotes, E. K.; Wheeler, E.; Soranzo, N.; Park, J. H.; Yang, J.; Gudbjartsson, D.; Heard-Costa, N. L.; Randall, J. C.; Qi, L.; Vernon Smith, A.; Mägi, R.; Pastinen, T.; Liang, L.; Heid, I. M.; Luan, J.; Thorleifsson, G.; Winkler, T. W.; Goddard, M. E.; Sin Lo, K.; Palmer, C.; Workalemahu, T.; Aulchenko, Y. S.; Johansson, A.; Zillikens, M. C.; Feitosa, M. F.; Esko, T.; Johnson, T.; Ketkar, S.; Kraft, P.; Mangino, M.; Prokopenko, I.; Absher, D.; Albrecht, E.; Ernst, F.; Glazer, N. L.; Hayward, C.; Hottenga, J. J.; Jacobs, K. B.; Knowles, J. W.; Kutalik, Z.; Monda, K. L.; Polasek, O.; Preuss, M.; Rayner, N. W.; Robertson, N. R.; Steinthorsdottir, V.; Tyrer, J. P.; Voight, B. F.; Wiklund, F.; Xu, J.; Zhao, J. H.; Nyholt, D. R.; Pellikka, N.; Perola, M.; Perry, J. R.; Surakka, I.; Tammesoo, M. L.; Altmaier, E. L.; Amin, N.; Aspelund, T.; Bhangale, T.; Boucher, G.; Chasman, D. I.; Chen, C.; Coin, L.; Cooper, M. N.; Dixon, A. L.; Gibson, Q.; Grundberg, E.; Hao, K.; Juhani Junttila, M.; Kaplan, L. M.; Kettunen, J.; König, I. R.; Kwan, T.; Lawrence, R. W.; Levinson, D. F.; Lorentzon, M.; McKnight, B.; Morris, A. P.; Müller, M.; Suh Ngwa, J.; Purcell, S.; Rafelt, S.; Salem, R. M.; Salvi, E.; Sanna, S.; Shi, J.; Sovio, U.; Thompson, J. R.; Turchin, M. C.; Vandenput, L.; Verlaan, D. J.; Vitart, V.; White, C. C.; Ziegler, A.; Almgren, P.; Balmforth, A. J.; Campbell, H.; Citterio, L.; De Grandi, A.; Dominiczak,

A.; Duan, J.; Elliott, P.; Elosua, R.; Eriksson, J. G.; Freimer, N. B.; Geus, E. J.; Glorioso, N.; Haiqing, S.; Hartikainen, A. L.; Havulinna, A. S.; Hicks, A. A.; Hui, J.; Igl, W.; Illig, T.; Jula, A.; Kajantie, E.; Kilpeläinen, T. O.; Koiranen, M.; Kolcic, I.; Koskinen, S.; Kovacs, P.; Laitinen, J.; Liu, J.; Lokki, M. L.; Marusic, A.; Maschio, A.; Meitinger, T.; Mulas, A.; Paré, G.; Parker, A. N.; Peden, J. F.; Petersmann, A.; Pichler, I.; Pietiläinen, K. H.; Pouta, A.; Ridderstråle, M.; Rotter, J. I.; Sambrook, J. G.; Sanders, A. R.; Schmidt, C. O.; Sinisalo, J.; Smit, J. H.; Stringham, H. M.; Bragi Walters, G.; Widen, E.; Wild, S. H.; Willemsen, G.; Zagato, L.; Zgaga, L.; Zitting, P.; Alavere, H.; Farrall, M.; McArdle, W. L.; Nelis, M.; Peters, M. J.; Ripatti, S.; van Meurs, J.B.; Aben, K. K.; Ardlie, K. G.; Beckmann, J. S.; Beilby, J. P.; Bergman, R. N.; Bergmann, S.; Collins, F. S.; Cusi, D.; den Heijer, M.; Eiriksdottir, G.; Gejman, P. V.; Hall, A. S.; Hamsten, A.; Huikuri, H. V.; Iribarren, C.; Kähönen, M.; Kaprio, J.; Kathiresan, S.; Kiemeney, L.; Kocher, T.; Launer, L. J.; Lehtimäki, T.; Melander, O.; Mosley, T. H.; Musk, A. W.; Nieminen, M. S.; O'Donnell, C. J.; Ohlsson, C.; Oostra, B.; Palmer, L. J.; Raitakari, O.; Ridker, P. M.; Rioux, J. D.; Rissanen, A.; Rivolta, C.; Schunkert, H.; Shuldiner, A. R.; Siscovick, D. S.; Stumvoll, M.; Tönjes, A.; Tuomilehto, J.; van, O. m. m. e. n.; Viikari, J.; Heath, A. C.; Martin, N. G.; Montgomery, G. W.; Province, M. A.; Kayser, M.; Arnold, A. M.; Atwood, L. D.; Boerwinkle, E.; Chanock, S. J.; Deloukas, P.; Gieger, C.; Grönberg, H.; Hall, P.; Hattersley, A. T.; Hengstenberg, C.; Hoffman, W.; Lathrop, G. M.; Salomaa, V.; Schreiber, S.; Uda, M.; Waterworth, D.; Wright, A. F.; Assimes, T. L.; Barroso, I.; Hofman, A.; Mohlke, K. L.; Boomsma, D. I.; Caulfield, M. J.; Cupples, L. A.; Erdmann, J.; Fox, C. S.; Gudnason, V.; Gyllensten, U.; Harris, T. B.; Hayes, R. B.; Jarvelin, M. R.; Mooser, V.; Munroe, P. B.; Ouwehand, W. H.; Penninx, B. W.; Pramstaller, P. P.; Quertermous, T.; Rudan, I.; Samani, N. J.; Spector, T. D.; Völzke, H.; Watkins, H.; Wilson, J. F.; Groop, L. C.; Haritunians, T.; Hu, F. B.; Kaplan, R. C.; Metspalu, A.; North, K. E.; Schlessinger, D.; Wareham, N. J.; Hunter, D. J.; O'Connell, J. R.; Strachan, D. P.; Wichmann, H. E.; Borecki, I. B.; van, D. u. i. j. n.; Schadt, E. E.; Thorsteinsdottir, U.; Peltonen, L.; Uitterlinden, A. G.; Visscher, P. M.; Chatterjee, N.; Loos, R. J.; Boehnke, M.; McCarthy, M. I.; Ingelsson, E.; Lindgren, C. M.; Abecasis, G. R.; Stefansson, K.; Frayling, T. M.; Hirschhorn, J. N. Hundreds of variants clustered in genomic loci and biological pathways affect human height. **Nature**, v. 467, p. 832-838, 2010.

Lettre, G.; Jackson, A. U.; Gieger, C.; Schumacher, F. R.; Berndt, S. I.; Sanna, S.; Eyheramendy, S.; Voight, B. F.; Butler, J. L.; Guiducci, C.; Illig, T.; Hackett, R.; Heid, I. M.; Jacobs, K. B.; Lyssenko, V.; Uda, M. ; Diabetes Genetics Initiative; FUSION; KORA; Prostate, Lung Colorectal and Ovarian Cancer Screening Trial; Nurses' Health Study; SardNIA; Boehnke, M.; Chanock, S. J.; Groop, L. C.; Hu, F. B.; Isomaa, B.; Kraft, P.; Peltonen, L.; Salomaa, V.; Schlessinger, D.; Hunter, D. J.; Hayes, R. B.; Abecasis, G. R.; Wichmann, H. E.; Mohlke, K. L.; Hirschhorn, J. N.

Identification of ten loci associated withheight highlights new biological pathways in human growth. **Nature Genetics**, v. 40, p. 584-591, 2008.

Lindholm-Perry, A. K.; Kuehn, L. A.; Smith, T. P.; Ferrell, C. L.; Jenkins, T. G.; Freetly, H. C.; Snelling, W. M. A region on BTA14 that includes the positional candidate genes LYPLA1, XKR4 and TMEM68 is associated with feed intake and growth phenotypes in cattle. **Animal Genetics**, v. 43, n. 2, p. 216-219, 2012.

Littlejohn, M.; Grala, T.; Sanders, K.; Walker, C.; Waghorn, G.; Macdonald, K.; Coppieters, W.; Georges, M.; Spelman, R.; Hillerton, E.; Davisand, S.; Snell, R. Genetic variation in PLAG1 associates with early life body weight and peripubertal weight and growth in Bos taurus. **Animal Genetics**, v. 43, n. 5, p. 591-594, 2012.

Maltecca, C.; Weigel, K. A.; Khatib, H.; Cowan, M.; Bagnato, A Whole-genome scan for quantitative trait loci associated with birth weight, gestation length and passive immune transfer in a Holstein x Jersey crossbred population. **Animal Genetics**, v. 40, n. 1, p. 27-34, 2009.

Mardia, K. V. Some properties of classical multi-dimensional scaling. **Communications on Statistics-Theory and Methods**, v. A7, n. 13, p. 1233-1241, 1978.

Matukumalli, L. K.; Lawley, C. T.; Schnabel, R. D.; Taylor, J. F.; Allan, M. F.; Heaton, M. P.; O'Connell, J.; Moore, S. S.; Smith, T. P.; Sonstegard, T. S.; Van Tassell, C. P. Development and characterization of a high density SNP genotyping assay for cattle. **PLoS One**, 4:e5350, 2009.

McClure, M. C.; Morsci, N. S.; Schnabel, R. D.; Kim, J. W.; Yao, P.; Rolf, M. M.; McKay, S. D.; Gregg, S. J.; Chapple, R. H.; Northcutt, S. L.; Taylor, J. F. A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. **Animal Genetics**, v. 41, n. 6, p. 597-607, 2010.

Meyer, K. Estimates of genetic parameters for mature weight of Australian beef cows and its relationship to early growth and skeletal measures. **Livestock Production Science**, v. 44, p. 125-137, 1995.

Nishimura, S.; Watanabe, T.; Mizoshita, K.; Tatsuda, K.; Fujita, T.; Watanabe, N.; Sugimoto, Y.; Takasuga, A.. Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. **BMC Genetics**, 13:40, 2012.

Nobre, P. R.; Misztal, I.; Tsuruta, S.; Bertrand, J. K.; Silva, L. O.; Lopes, O. S. Analyses of growth curves of Nellore cattle by multiple-trait and random regression models. **Journal of Animal Science**, v. 81, p. 918-926, 2003.

Pausch, H.; Flisikowski, K.; Jung, S.; Emmerling, R.; Edel, C.; Götz, K. U.; Fries, R. Genome-wide association study identifies two major loci affecting calving ease and growth-related traits in cattle. **Genetics**, v. 187, p. 289-297, 2011.

Phocas, F.; Bloch, C.; Chapelle, P.; Bécherel, F.; Renand, G.; Ménissier, F. Developing a breeding objective for a French purebred beef cattle selection programme. **Livestock Production Science**, v. 57, p. 49-65, 1998.

Porto-Neto, L. R.; Bunch, R. J.; Harrison, B. E.; Barendse, W. Variation in the XKR4 gene was significantly associated with subcutaneous rump fat thickness in indicine and composite cattle. **Animal Genetics**, v. 43, n. 6, p. 785-789, 2012.

Pryce, J.E.; Hayes, B.J.; Bolormaa, S.; Goddard, M.E. Polymorphic regions affecting human height also control stature in cattle. **Genetics**, v. 187, p. 981-984, 2011.

R Development Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing 2008, Vienna, Austria**. Available at: <http://www.R-project.org>, Accessed on 19 dec. 2012.

Snelling, W. M.; Allan, M. F.; Keele, J. W.; Kuehn, L. A.; McDaneld, T.; Smith, T. P.; Sonstegard, T. S.; Thallman, R. M.; Bennett, G. L. Genome-wide association study of growth in crossbred beef cattle. **Journal of Animal Science**, v. 88, n. 3, p. 837-848, 2010.

Spelman, R. J.; Huisman, A. E.; Singireddy, S. R.; Coppieters, W.; Arranz, J.; Georges, M.; Garrick, D. J. Short communication: quantitative trait loci analysis on 17

nonproduction traits in the New Zealand dairy population. **Journal of Dairy Science**, v. 82, n. 11, p. 2514-2516, 1999.

The Bovine Genome Sequencing and Analysis Consortium; Elsik, C. G.; Tellam, R. L.; Worley, K. C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, v. 324, n. 5926, p. 522-528, 2009.

The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. **Science**, v. 324, n. 5926, p. 528-532, 2009.

Van Dyck, F.; Declercq, J.; Braem, C.V.; Van de Ven, W.J. PLAG1, the prototype of the PLAG gene family: versatility in tumour development. **International Journal of Oncology**, v. 30, p. 765-774, 2007.

Weedon, M. N.; Lango, H.; Lindgren, C. M.; Wallace, C.; Evans, D. M.; Mangino, M.; Freathy, R. M.; Perry, J. R.; Stevens, S.; Hall, A. S.; Samani, N. J.; Shields, B.; Prokopenko, I.; Farrall, M.; Dominiczak, A. ; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium; Johnson, T.; Bergmann, S.; Beckmann, J. S.; Vollenweider, P.; Waterworth, D. M.; Mooser, V.; Palmer, C. N.; Morris, A. D.; Ouwehand, W. H. ; Cambridge GEM Consortium; Zhao, J. H.; Li, S.; Loos, R. J.; Barroso, I.; Deloukas, P.; Sandhu, M. S.; Wheeler, E.; Soranzo, N.; Inouye, M.; Wareham, N. J.; Caulfield, M.; Munroe, P. B.; Hattersley, A. T.; McCarthy, M. I.; Frayling, T. M. Genome-wide association analysis identifies 20 loci that influence adult height. **Nature Genetics**, v. 40, p. 575-583, 2008.

CHAPTER 3 - Genome-wide mapping of loci explaining variance in scrotal circumference in Nellore cattle

Utsunomiya, Y. T.; Carmo, A. S.; Neves, H. H. R.; Carvalheiro, R. Matos, M. C.; Zavarez, L. B.; Ito, P. K. R. K.; Pérez O'Brien, A. M.; Sölkner, J.; Porto-Neto, L. R.; Schenkel, F. S.; McEwan, J. C.; Cole, J. B.; da Silva, M. V. G. B.; Van Tassell, C. P.; Sonstegard, T. S.; Garcia J. F.

PLoS ONE (Accepted for publication)

1. Abstract

The reproductive performance of bulls has a high impact on the beef cattle industry. Scrotal circumference (SC) is the most recorded reproductive trait in beef herds, and is used as a major selection criterion to improve precocity and fertility. The characterization of genomic regions affecting SC can contribute to the identification of diagnostic markers for reproductive performance and uncover molecular mechanisms underlying complex aspects of bovine reproductive biology. In this paper, we report a genome-wide scan for chromosome segments explaining differences in SC, using data of 861 Nellore bulls (*Bos indicus*) genotyped for over 777,000 single nucleotide polymorphisms. Loci that excel from the genome background were identified on chromosomes 4, 6, 7, 10, 14, 18 and 21. The majority of these regions were previously found to be associated with reproductive and body size traits in cattle. The signal on chromosome 14 replicates the pleiotropic quantitative trait locus encompassing *PLAG1* that affects male fertility in cattle and stature in several species. Based on intensive literature mining, *SP4*, *MAGEL2*, *SH3RF2*, *PDE5A* and *SNAI2* are proposed as novel candidate genes for SC, as they affect growth and testicular size in other animal models. These findings contribute to linking reproductive phenotypes to gene functions, and may offer new insights on the molecular biology of male fertility.

Keywords: *Bos indicus*, testicular size, male fertility, SNP

2. Introduction

Reproductive performance has a high economic value in beef cattle, because fertility affects generation intervals, the intensity of selection pressure that can be applied to the population, and the amount of product that can be sent to the market (VAN MELIS et al., 2010). Furthermore, reproductive wastage is a major reason for culling beef cows.

Domestic cattle are composed by two interfertile species: the humpless taurine cattle (*Bos taurus*) and the humped indicine or zebu cattle (*Bos indicus*). Indicine breeds, such as Nellore cattle, form the majority of the beef herds in tropical and subtropical countries. Zebus generally take longer to reach puberty than taurines (MARTIN et al., 1992), making the improvement of reproductive performance an impending challenge in the production systems of these regions of the world.

Scrotal circumference evaluated at yearling (SC) is the most recorded reproductive trait in breeding programs for beef cattle, as the trait is inexpensive and easy to measure (BALL et al., 1983), is highly heritable (DIAS et al., 2003), and is associated with testis development, quantitative and qualitative semen parameters (BOURDON & BRINKS, 1986), age at puberty in bulls and related heifers (TOELLE & ROBISON, 1985; EVANS et al., 1999), heifer pregnancy (VAN MELIS et al., 2010), and body weight (BERGMANN et al., 1996). Consequently, SC is used in these programs as a major indicator of precocity and fertility.

Characterizing genomic regions that explain differences in SC in *B. indicus* can contribute to the identification of reproductive performance informative molecular markers to assist breeding, as well as to the mapping of loci implicated in reproductive biology. In this paper, we analyzed data of estimated breeding values (EBV) from 861 Nellore bulls genotyped for over 777,000 single nucleotide polymorphism (SNP) markers. We aimed at identifying putative genomic regions explaining differences in SC in *B. indicus* cattle via genome-wide mapping.

3. Material and methods

3.1. Ethical statement

Local ethical committee approval was not necessary in the present study, because phenotypic data were obtained from a database (ALIANÇA, 2012), and DNA samples used for genotyping were obtained from commercialized semen straws.

3.2. Animals and genotypes

Estimated breeding values for SC were obtained from routine genetic evaluations (ALIANÇA, 2012), comprising data from 542,918 animals born between 1985 and 2011, and raised in 243 grazing-based Brazilian herds. Scrotal circumference in yearlings (around 18 months of age) was measured as recommended by the *Society for Theriogenology* (BALL et al., 1983). Genetic evaluation of SC was based on two single-trait animal models, both including a fixed effect of contemporary group (defined as animals from the same herd, born in the same year and season, and belonging to the same management group from birth until yearling), and random effects that included direct additive genetic, maternal additive genetic, maternal permanent environmental and residual error effects. In the first model (SC_A), the fixed effect of age at SC measurement was included as a covariate. In the second model (SC_{AW}), covariates accounting for differences due to age and weight at yearling were included. Estimated breeding values were deregressed by the method described by Garrick et al. (2009) and treated as pseudo-phenotypes in the genome-wide mapping analysis.

3.3. Genotyping and data filtering

Only sires widely used via artificial insemination whose accuracies (i.e., square root of reliability, calculated based on prediction error variance estimates) for SC_A and SC_{AW} were greater than 0.5 were considered for genotyping. A total of 861

progeny-tested Nellore bulls were genotyped for 777,962 SNPs with the Illumina® BovineHD Genotyping BeadChip assay, according to the manufacturer's protocol. As a first filtering criterion, only samples with call rate greater than 0.9 and SNPs with GenTrain score greater than or equal to 0.7 were considered for analysis. Mitochondrial DNA and unmapped markers were also excluded.

As males are hemizygous for both sex chromosomes, observation of heterozygous X and Y genotypes are only possible for SNP probes that hybridize against the XY pseudo-autosomal region (PAR). As the *UMD v3.1* bovine genome assembly (ZIMIN et al., 2009) does not allow for clear distinction of PAR markers, all heterozygous X- and Y-linked genotypes were considered as genotyping errors and set to missing. Next, SNPs were removed from the dataset if they did not exhibit minor allele frequency greater than or equal to 0.02 or call rate of at least 0.98. These procedures and many others described later were performed using customized functions and the *base* and the *GenABEL v1.7-6* packages in *R v2.15.0* (AULCHENKO et al., 2007; R DEVELOPMENT CORE TEAM, 2008).

3.4. Genome-wide mapping

We adapted the two-steps *Fast Association Score Test-based Analysis* (FASTA) (CHEN & ABECASIS, 2007) to compute allele substitution effects accounting for relatedness, population structure and heterogeneity of variance in deregressed EBVs (dEBVs). In the first step, the variance-covariance matrix for the pseudo-phenotypes was estimated using an animal model that included random additive genetic and residual effects. In the second step, the estimated variance-covariance matrix was used to compute allele substitution effects for each SNP via generalized least squares. A detailed description of this analysis can be found in Appendix A.

Next, aiming at mapping loci explaining differences in SC, we investigated chromosome windows where the average phenotypic variance explained by SNPs deviated substantially from the genome background. First, the percentage of phenotypic variance explained by each SNP was calculated as:

$$\%\hat{\pi}_i = \frac{2p_i q_i \hat{\beta}_i^{\,2}}{\sigma_T^2} \text{x } 100$$

where, relative to SNP i, $\hat{\beta}_i$ is the estimated allele substitution effect, p_i and q_i are the allele frequencies, and σ_T^2 is the total trait variance.

Second, in order to reduce noise and improve mapping, the phenotypic variance explained by SNPs was smoothed across the genome by averaging $\%\hat{\pi}_i$ in sliding windows of 1 Mb, sliding 50 kb at a time. Only windows containing at least 10 SNPs were averaged, and we considered as outliers the windows where $\%\hat{\pi} > 5 \text{ x } IQR + Q_3$, where IQR is the interquartile range and Q_3 is the third quartile of the distribution. Third, we used *BEDTools v2.12.0* (QUINLAN & HALL, 2010) to merge the intervals of overlapping outlier windows. These merged windows were considered as candidate loci for SC.

We assessed the relative mapping resolution gain when using sliding windows instead of single SNPs by calculating the signal-to-noise ratio between the two strategies. For each approach, we let the signal-to-noise coefficient be represented by the reciprocal of the coefficient of variation $SN = \mu/\sigma$, where μ and σ are the mean percentage of phenotypic variance explained and its standard deviation, respectively, by either single SNPs or windows. Then, we calculated the ratio of the signal-to-noise coefficients between strategies as SN_{window}/SN_{SNP}.

3.5. Assessment of functional relevance

The *cattle QTLdb* database (HU et al., 2013) was examined to find out if any genomic region identified here overlapped with a previously described bovine quantitative trait locus (QTL), in particular those related to body size and reproductive traits. Gene coordinates in the *UMD v3.1* assembly (ZIMIN et al., 2009) were obtained from the *Ensembl genes 73* database using the *BioMart tool* (KINSELLA et al., 2011), and overlaps between the boundaries of candidate loci and gene coordinates were determined using *BEDTools v2.12.0* (QUINLAN & HALL, 2010). Finally, we conducted intensive literature review to propose functionally sound candidate genes associated with SC.

4. Results

A total of 525,961 SNPs (67.6%) and 861 individuals (100.0%) passed the filtering criteria and were retained in the dataset. After filtering, the average and median gap size between consecutive markers were 5.1 kb and 3.3 kb, respectively, indicating high density coverage of the genome. In spite of the effort and interest to analyze both sex chromosomes, Y SNPs did not present sufficient variability to allow for estimation of allele substitution effects.

For both SC_A and SC_{Aw}, the distributions of pseudo-phenotypes were approximately normal, and the dEBVs for these two traits were fairly correlated, with R^2 = 0.74 (Figure 1). Accuracies were virtually equal for SC_A and SC_{Aw} (identical up to the second decimal place), with mean, minimum, median and maximum of 0.84 ± 0.11, 0.51, 0.86 and 0.98, respectively.

A total of 52,493 SNP windows of 1 Mb were built across the genome, with an average density of 198 ± 59 SNPs per window. The ratio between the signal-to-noise coefficients of sliding windows and single SNPs was 2.54, indicating that the smoothing strategy allowed for a 2.54 fold improvement in mapping resolution (Figure 2). Considering the $5 \times IQR + Q_3$ threshold for percentage of phenotypic variance explained (SC_A = 0.40% and SC_{Aw} = 0.42%), 236 (0.45%) and 279 (0.53%) windows were declared outliers for SC_A and SC_{AW}, respectively. After merging overlapping outlier windows, we obtained a total of 8 and 6 candidate loci explaining approximately 4% of the variance in SC_A and SC_{AW}, respectively (Table 1).

Overall, results from the genome-wide mapping analysis were strikingly similar between SC_A and SC_{AW}, indicating that fitting the covariate for weight at yearling in the model did not cause substantial mapping differences. Indeed, four loci were shared between the two traits in chromosomes 6, 10, 14 and 21, respectively (Figure 3), which exhibited clear signals in the genome-wide plot of smoothed phenotypic variance explained by SNP windows (Figure 2). From the total of six differentially detected loci, one was located nearby the shared locus on chromosome 14 (Table 1), and the three SC_{AW} private loci on chromosomes 10 and 7 were close to be declared outliers for SC_A (Figure 2).

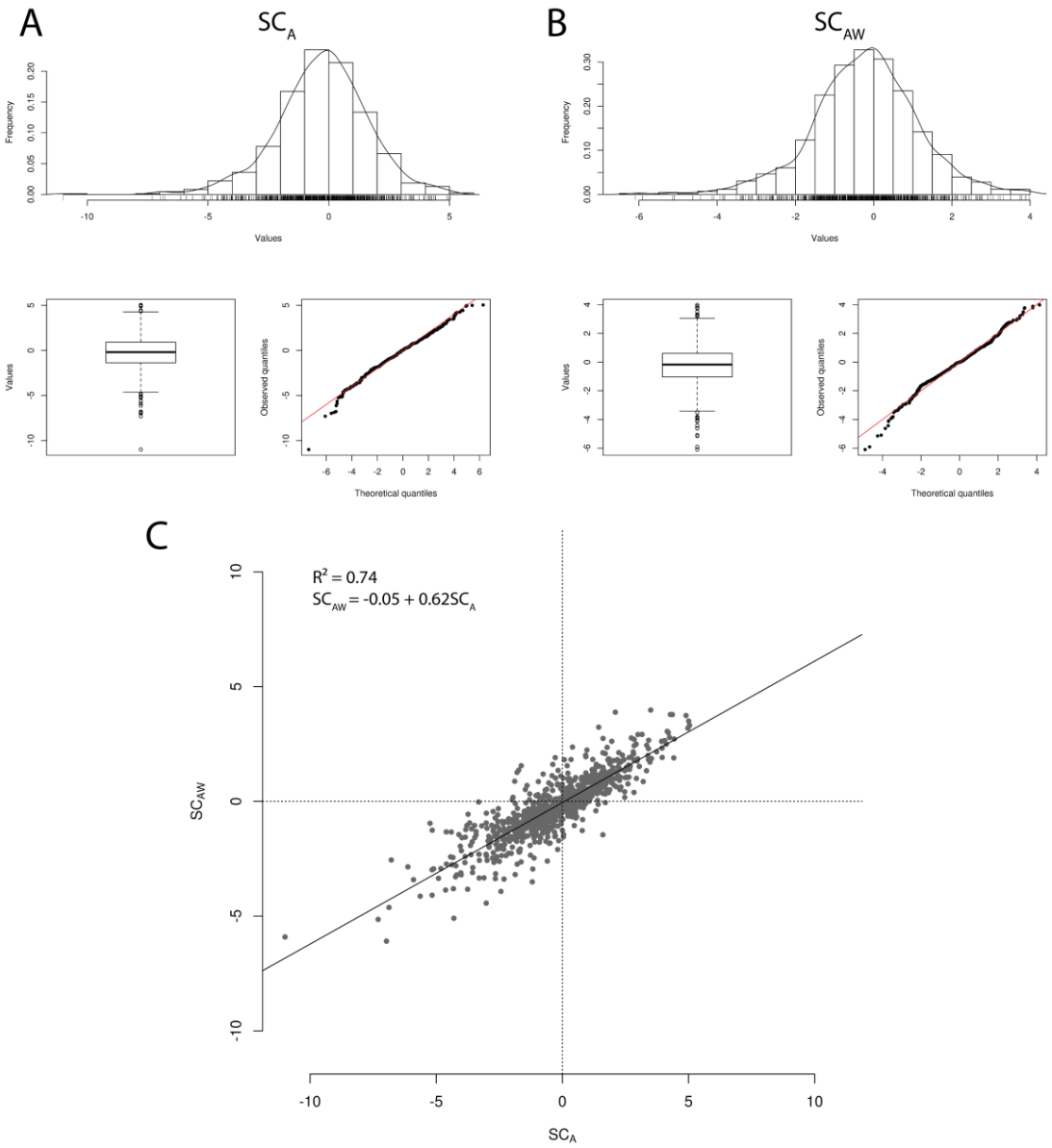

Figure 1. Descriptive statistics for scrotal circumference dEBVs of 861 Nellore bulls. Histograms (top), boxplot (bottom left) and normal quantile-quantile plots (bottom right) are provided for scrotal circumference A) corrected for age (SC_A) and B) corrected for age and weight at yearling (SC_{AW}). A scatter plot illustrating the linear relationship between the two dEBVs is also provided (C).

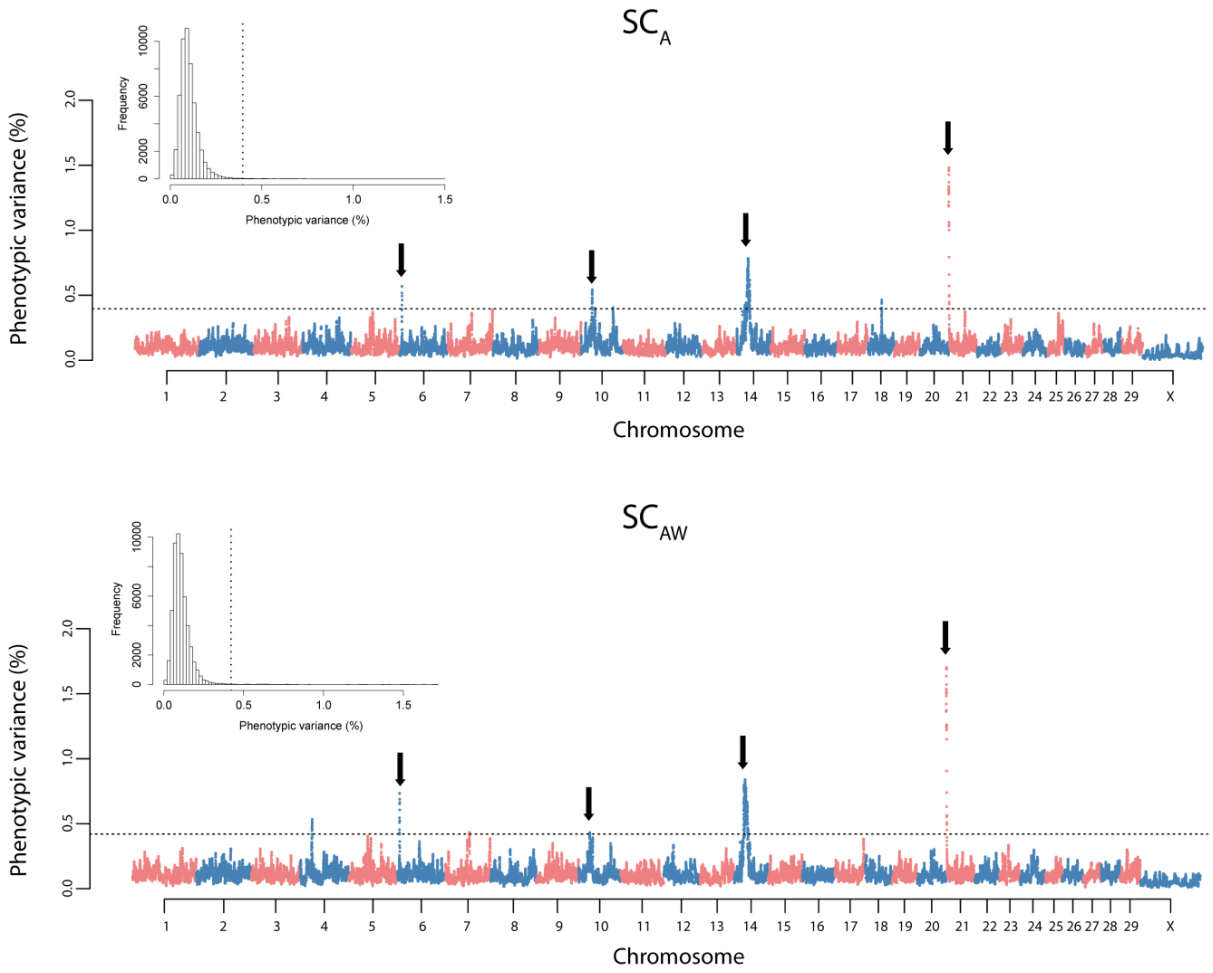

Figure 2. Manhattan plots of scrotal circumference variance explained by SNP windows in Nellore cattle. Pseudo-phenotypes were based on dEBVs corrected for age (SC_A) and corrected for age and weight at yearling (SC_{AW}). Each dot represents a 1 Mb SNP window. Horizontal dashed lines represent adopted thresholds (SC_A = 0.40% and SC_{AW} = 0.42%). Arrows indicate signals shared between the two models. Histograms represent the distribution of phenotypic variance explained by SNP windows, and the dotted vertical line marks the adopted thresholds.

Table 1. Detected major loci explaining variance in scrotal circumference in Nellore cattle.

Scrotal circumference model	Chromosome	Position start (Mb)	Position end (Mb)	Peak position (Mb)	Segment length (Mb)	Number of SNPs	Average MAF[a]	Average $\%\hat{\pi}$ [b]	Functional candidate gene
Corrected for age (SC_A)	6	5.20	6.65	5.80-6.00	1.45	42	0.28	0.52	*PDE5A (ENSBTAG00000024888)*
	10	26.90	28.80	28.00	1.90	387	0.23	0.49	*C15ORF55* (ENSBTAG00000014948)
	10	34.80	35.80	35.30	1.00	140	0.19	0.40	*FSIP1 (ENSBTAG00000012015)*
	10	78.85	79.85	79.35	1.00	122	0.20	0.41	-
	14	20.25	21.45	20.90	1.20	310	0.24	0.41	*SNAI2* (ENSBTAG00000013227)
	14	23.40	33.85	29.10	10.45	2236	0.23	0.55	*PLAG1* (ENSBTAG00000004022)
	18	34.55	35.80	35.20	1.25	206	0.22	0.44	*CES4A* (ENSBTAG00000038325)
	21	0.00	2.50	1.20	2.50	127	0.17	1.07	*MAGEL2* (ENSBTAG00000045998)
Corrected for age and weight at yearling (SC_{AW})	4	28.95	30.80	30.15	1.85	301	0.20	0.48	*SP4* (ENSBTAG00000014389)
	6	5.05	6.75	5.80-6.00	1.70	84	0.27	0.63	*PDE5A (ENSBTAG00000024888)*
	7	59.15	60.25	59.75	1.10	158	0.16	0.43	*SH3RF2* (ENSBTAG00000006762)
	10	27.15	28.25	27.70	1.10	209	0.21	0.43	*C15ORF55* (ENSBTAG00000014948)
	14	23.25	35.50	25.00 and 27.00	12.25	2687	0.23	0.63	*PLAG1* (ENSBTAG00000004022)
	21	0.00	2.50	1.20	2.50	127	0.17	1.23	*MAGEL2* (ENSBTAG00000045998)

[a] MAF = Minor allele frequency.

[b] $\%\hat{\pi}$ = Average percentage of phenotypic variance explained by overlapping 1 Mb SNP windows

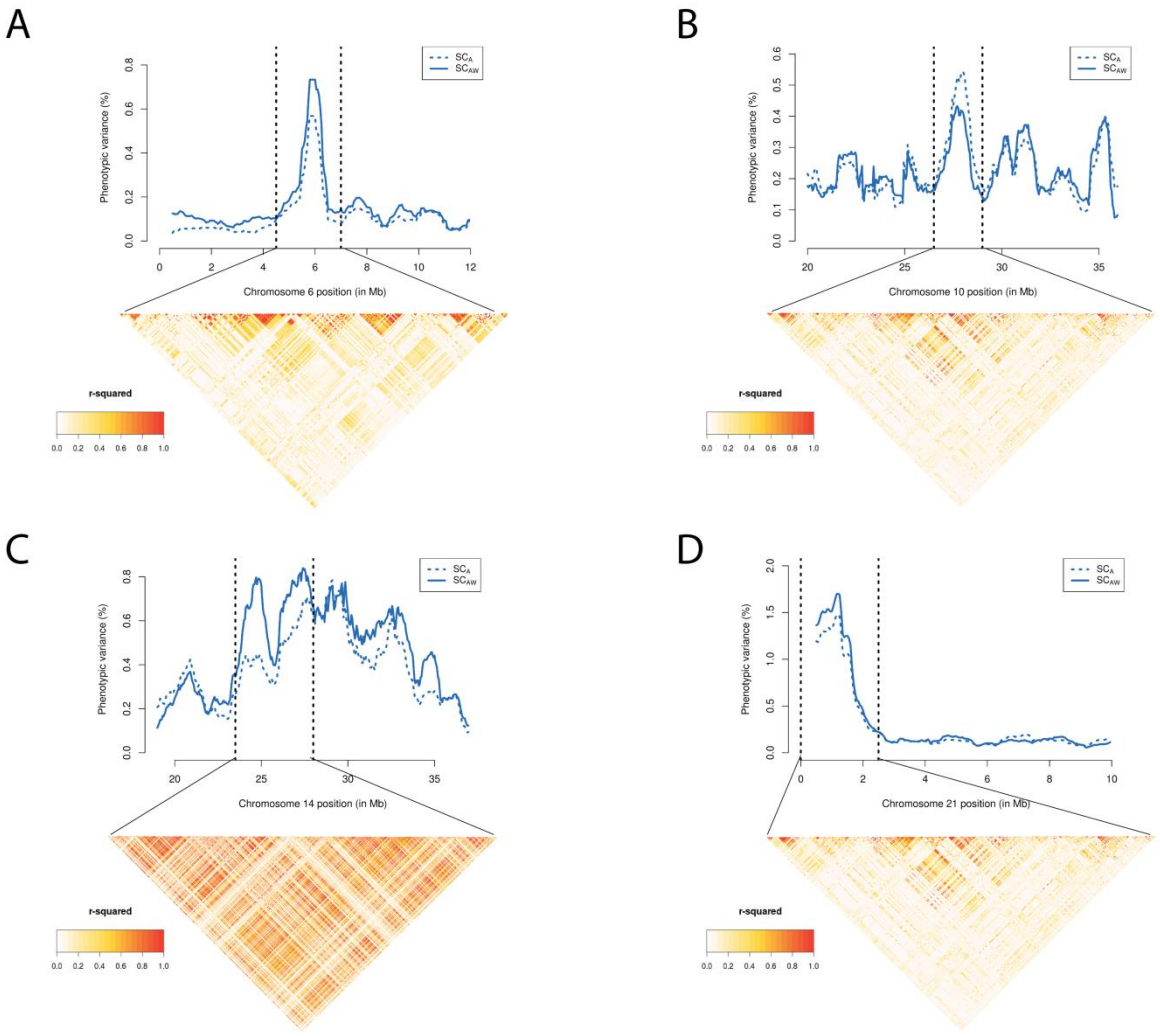

Figure 3. Regional plots of scrotal circumference variance explained by SNP windows in Nellore cattle. Pseudo-phenotypes were based on dEBVs corrected for age (SC_A) and corrected for age and weight at yearling (SC_{AW}). Clear common signals between SC_A and SC_{AW} were found on chromosomes A) 6, B) 10, C) 14 and D) 21. Vertical black dashed lines delimit the regions where the highest variance explained were found. Linkage disequilibrium structure for these regions (bottom) is portrayed as a heatmap of r^2 values between SNPs.

A total of 285 and 190 genes were mapped against the major loci found for SC_A and SC_{AW}, respectively, and a total of 309 unique genes were observed. From these, 246 protein coding, 25 snoRNA, 12 snRNA, 9 miRNA, 9 rRNA, 1 misc_RNA

and 7 pseudo genes were observed. From this gene list, we filtered 9 functional candidates implicated in growth, testicular size and fertility (Table 1), which included: pleiomorphic adenoma gene 1 (*PLAG1*, ENSBTAG- 00000004022), carboxylesterase 4A (*CES4A*, ENSBTAG00000038325), Sp4 transcription factor (*SP4*, ENSBTAG00000014389), melanoma antigen family L 2 (*MAGEL2*, ENSBTAG00000045998), phosphodiesterase 5A (*PDE5A*, ENSBTAG00000024888), snail family zinc finger 2 (*SNAI2*, ENSBTAG- 00000013227), nuclear protein in testis (*C15ORF55*, ENSBTAG00000014948), fibrous sheath-interacting protein 1 (*FSIP1*, ENSBTAG00000012015), and SH3 domain containing ring finger 2 (*SH3RF2*, ENSBTAG00000006762).

A total of 76 production and reproduction QTLs, mined from 24 distinct publications, were mapped against the loci found here. The largest trait contingency observed was composed by traits related to body size (43 QTLs), followed by reproductive traits (23 QTLs). Furthermore, the locus detected on chromosome 4 (Table 1) mapped against one previously described QTL for SC in Angus cattle (*B. taurus*) (MCCLURE et al., 2010).

5. Discussion

The genome-wide mapping analysis detected positional candidate loci explaining approximately 4% of the dEBVs for SC (Table 1). Although this represents only a fraction of the trait variance, this percentage is substantial considering that 180 loci associated with human adult height, a highly heritable and classic polygenic trait, explain only 10% of the phenotypic variance together (LANGO et al., 2010). This is evidence that multiple loci across the genome are involved in the complex inheritance of SC, and the functional candidate genes filtered here may only scratch the surface of the molecular mechanisms underlying the trait. The dissection of the pathways regulating precocity in *B. indicus* cattle will require multiple studies across breeds and trait models, with intensive multidisciplinary reasoning. Nevertheless, the loci reported here excel from the genome background, and represent important data in the context of bovine reproductive biology.

The region explaining the largest proportion of SC variance in the present study mapped to the beginning of chromosome 21, peaking around 1.5 Mb. The closest gene found in this region was *MAGEL2*. The orthologous human and murine genes regulate normal circadian output, and are highly expressed in the suprachiasmatic nucleus of the hypothalamus (KOZLOV et al., 2007). The human *MAGEL2* has been implicated in Prader-Willi Syndrome, a genetic disorder characterized by short stature, low muscle tone, cognitive disabilities, increased food intake, obesity, low levels of insulin and insulin-like growth factor 1 (IGF1), incomplete sexual development, hypogonadism, and male infertility (KOZLOV et al., 2007; BISCHOF et al., 2007). The disorder manifests when a segment on human chromosome 15, which encompasses seven maternally imprinted genes including *MAGEL2*, presents a deletion or loss of expression of the paternal alleles. Inactivation of the mouse *MAGEL2* alone was shown to lead to abnormalities suggestive of hypothalamic dysfunction similar to the Prader-Willi Syndrome (BISCHOF et al., 2007).

Of note, variation of copy number gain spanning the interval between 1.57 Mb and 2.99 Mb on bovine chromosome 21 has been found by the comparison of individual whole genome sequence data of Nellore with the *B. taurus* breeds Angus, Holstein and Hereford (BICKHART et al., 2012). As the Prader-Willi Syndrome is caused by loss of the paternal copy of the orthologous sequence in humans, and *MAGEL2* is essential for proper hypothalamic control of growth and fertility (KOZLOV et al., 2007), association of copy number variants with growth and reproductive traits seems to be a sensible hypothesis to be tested on this chromosome segment.

The locus detected on chromosome 7 encompasses *SH3RF2*. Rubin et al. (2010) discovered a deletion removing all but the first exon of the orthologous chicken gene that is associated with body weight, and demonstrated that strong selection caused the deletion to reach fixation in a high growth lineage. Interestingly, using a mouse model of Prader-Willi syndrome, Stefan et al. (2005) found that loss of expression of the *MAGEL2* region induces upregulation of *SH3RF2* and its flanking genes *TCERG1*, *LARS*, *RBM27* and *GPR151*. As both the *MAGEL2* and the *SH3RF2* regions were flagged in the present study, a trans-acting regulatory mechanism involving the loci on chromosomes 7 and 21 found here is likely to

underlie SC variation. Hence, these signals are plausible candidates for weight and male fertility traits in Nellore cattle.

We identified a candidate locus on chromosome 14 with the highest percentage of phenotypic variance explained mapping to positions 25 Mb and 27 Mb. Fortes *et al*. (2012b) reported associations for IGF1 at 6 months and SC at 12 months in young Brahman bulls (*B. indicus*) in an overlapping region around 25 Mb, which was previously shown to correlate with age of Brahman bulls when they achieve 26 cm of SC (FORTES et al., 2012a). This region has been well characterized in taurine cattle as harboring several human orthologues affecting stature and growth (PRYCE et al., 2011), especially *PLAG1* (KARIM et al., 2011). The locus has also been found to be associated with birth weight in Nellore cattle, and suggested to shelter polymorphisms with pleiotropic effects on traits that correlate with body size (UTSUNOMIYA et al., 2013). Furthermore, some first evidences for pleiotropism in body size and fertility traits in the *PLAG1* region have been recently found in Brahman cattle (FORTES et al., 2013).

Although the human stature orthologues flanking 25 Mb are appealing candidates for SC, the chromosome 14 signal found here comprises a large segment spanning from 20.25 Mb to 35.85 Mb. This may be evidence that multiple genes and variants within this region are involved. For instance, *SNAI2* is located around 21.58 Mb and encodes for Slug (also known as Slugh), a Zinc-finger transcription factor that when mutated in mice produces individuals with testicular atrophy and marked decrease in seminiferous tubules sizes (PÉREZ-LOSADA et al., 2002). Although these mice are able to copulate, their offspring are small. Also, Fortes *et al*. (2013b) showed that ability to produce sperm at 18 months in Brahman bulls is not as significant around 25 Mb, and exhibits signals of association shifted towards the 35 Mb position instead.

Another possible justification for a signal coming from such a large chromosome segment is a long range linkage disequilibrium (LD) persistency within the region. In fact, we found a strong LD structure underpinning the signal (Figure 3), which may be hampering the localization of the true locus involved. In either case, these evidences together support the entire chromosome segment identified here as a key region affecting growth and fertility traits in cattle.

The locus detected on chromosome 6 is located 124 kb downstream of *PDE5A*. The phosphodiesterase encoded by *PDE5A* is substantially expressed in the testis, and mice overexposed to inhibitors of this protein present testicular tissue alterations, including decreased testis weight, degeneration, and atrophy of the seminiferous epithelium (VEZZOSI & BERTHERAT, 2011). This genomic region also shelters genes that interact with other proteins previously linked to small testis size. For instance, the protein encoded by *MAD2* belongs to the mitotic checkpoint complex, and is recruited by the mitotic kinase Bub1. A residue change in the catalytic loop of Bub1 was shown to lead to male subfertility, with marked reduction in testicular size (RICKE et al., 2012).

Several QTLs mapping to the loci detected here were related either to body size or reproductive traits that are associated with SC. In particular, the peak on chromosome 4 mapped against one previously reported QTL for SC (MCCLURE et al., 2010), which encompasses *SP4*. This gene encodes for a zinc finger transcription factor that is predominantly expressed in the brain, but is also detectable in the testicular tissue (HAGEN et al., 1992; SUPP et al., 1996). Gölner *et al*. (2001) showed that *SP4*-knockout mice develop until birth without obvious abnormalities, but two-thirds of them die within 4 weeks after birth and the remaining one-third present growth retardation. Surviving male mice exhibit reduced testis size, although complete spermatogenesis can be observed. Surviving female mice exhibit small-sized thymus, spleen and uterus, and all mice show pronounced delay in sexual maturation. As *SP4*-knockout mice present growth retardation mainly after birth, it is likely that variations in the bovine *SP4* affect body size and testicular growth from birth to yearling age, but they are unlikely to affect fetal development or spermatogenesis. Moreover, the evidence of delayed sexual maturation and reduced testicular size in surviving *SP4*-knockout mice is consistent with the known positive correlation between SC and precocity in cattle.

The functional candidate gene surrounding the peak on chromosome 18 was *CES4A*, a hydrolase member of the carboxylesterase large family (enzyme class EC 3.1.1.1.-), also known as *CES6*, *CES8* or Hydrolase A. Carboxylesterases act in the transesterification of a broad spectrum of substrates, and play an important role in the metabolism of endogenous lipid and foreign compounds such as drugs and

pesticides (SATOH & HOSOKAWA, 2998). Hydrolase A is known to be expressed in several tissues, including the testis (YAN et al., 1995). Esterase activity (EC 3.1.1.-) has been found to be abundant in the testis and associated with androgen production (HUGGINS & MOULTON, 1948). Two intronic SNPs in the human *CES4A* were found to be correlated with high density lipoprotein levels (HDL) in an association analysis deposited in the *dbGaP* database (www.ncbi.nlm.nih.gov/gap, accession: pha002900.1, accessed on 21 Oct 2013), conducted in an expanded population sample from the original 1966 Northern Finland Birth Cohort (NFBC66) study (SABATTI et al., 2009). Also, testosterone treatment of aging men with hypogonadism was demonstrated to lower HDL levels, and the mechanism underlying this relationship and its role in coronary disease risk have been targets of debate and controversy (LANGER et al., 2002). These evidences together point to *CES4A* as a functional candidate gene affecting differences in SC in Nellore cattle, and the underlying mechanism may involve the dynamics between HDL and androgen levels.

Some of the genes lying within the loci detected nearby positions 28 Mb and 35 Mb on chromosome 10 are related to the testicular tissue, but their association with scrotal circumference is unclear. The nuclear protein in testis gene (*C15ORF55* or *NUTM1*) is mainly known by its involvement with midline organs carcinoma (FRENCH et al., 2003). The fibrous sheath-interacting protein 1 (*FSIP1*) was shown to bind to Akap4 during spermatogenesis in order to form the fibrous sheath of the sperm flagellum (BROWN et al., 2003). The peak found around 79.35 Mb on chromosome 10 did not reveal an appealing functional candidate gene, but the region overlaps QTLs for dystocia, fetal death and birth weight. Further investigation of these loci is needed to elucidate if they contribute to phenotypic variation in SC, as well as to clarify the molecular mechanisms underlying this contribution.

6. Conclusions

In summary, this is believed to be the first study applying a high-density SNP panel in a genome-wide survey of loci affecting scrotal circumference in Nellore cattle, which contributes with important preliminary data to the dissection of molecular

mechanisms regulating precocity in the bovine species. The loci identified here harbor known and novel functional candidate genes affecting scrotal circumference in *B. indicus* cattle. Fine mapping of these signals with whole genome sequence data and hypothesis-driven experiments may shed light on the genes and networks underlying phenotypic variation in fertility traits in cattle. In a broader perspective, as the majority of the genes found in eutherian mammals are orthologous, further investigation of these loci in cattle may offer new insights on complex aspects of mammalian reproductive biology.

7. Acknowledgments

We thank Guilherme Penteado Coelho Filho and Daniel Biluca for technical assistance in samples acquisition. Mention of trade name proprietary product or specified equipment in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the authors or their respective institutions.

8. Competing interest

The authors have declared that no competing interests exist.

9. Financial disclosure

This research was supported by: National Counsel of Technological and Scientific Development (CNPq - http://www.cnpq.br/) (process 560922/2010-8 and 483590/2010-0); and São Paulo Research Foundation (FAPESP - http://www.fapesp.br/) (process 2011/16643-2 and 2010/52030-2). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

10. Author’s contributions

Conceived and designed the experiments: J. F. Garcia, T. S. Sonstegard, J. Sölkner, J. McEwan, J. B. Cole, C. P. Van Tassell, F. S. Schenckel, M. V. G. B. Silva, L. R. Porto-Neto. Performed the experiments: A. S. Carmo, Y. T. Utsunomiya, H. H. R. Neves, R. Carvalheiro, M. C. Matos, L. B. Zavarez, P. K. R. K. Ito, A. M. Pérez O'Brien. Analyzed the data: Y. T. Utsunomiya. Contributed reagents/materials/analysis tools: J. F. Garcia, T. S. Sonstegard, J. Sölkner, J. McEwan, J. B. Cole, C. P. Van Tassell, F. S. Schenckel, M. V. G. B. Silva, L. R. Porto-Neto, R. Carvalheiro, H. H. R. Neves, A. S. Carmo, Y. T. Utsunomiya. Wrote the manuscript: Y. T. Utsunomiya. Coordinated the study: J. F. Garcia. All authors read, approved and contributed to edit the final manuscript.

11. References

Aliança. Sumário de touros 2012 Aliança Nelore. **GenSys Consultores Associados**, 2012. Available at: <http://www.gensys.com.br/home/show_page.php?id=701>, Accessed on 14 oct. 2013.

Aulchenko, Y. S.; Ripke, S.; Isaacs, A.; Van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. **Bioinformatics**, v. 23, n. 10, p. 1294-1296, 2007.

Ball, L.; Ott, R. S.; Mortimer, R.G.; Simons, J.C. Manual for Breeding Soundness Examination of Bulls. **Journal of the Society for Theriogenology**, v. 12, 65 p., 1983.

Bergmann, J. A. G.; Zamborlini L. C.; Procopio, C. S. O.; Andrade, V. J.; Vale Filho, V. R. Estimativas de parâmetros genéticos do perímetro escrotal e do peso corporal em animais da raça Nelore. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 48, p. 69-78, 1996.

Bickhart, D. M.; Hou, Y.; Schroeder, S. G.; Alkan, C.; Cardone, M. F.; Matukumalli, L. K.; Song, J.; Schnabel, R. D.; Ventura, M.; Taylor, J. F.; Garcia, J. F.; Van Tassell, C. P.; Sonstegard, T. S.; Eichler, E. E.; Liu, G. E. Copy number variation of individual cattle genomes using next-generation sequencing. **Genome Research**, v. 22, p. 778-790, 2012.

Bischof, J. M.; Stewart, C. L.; Wevrick, R. Inactivation of the mouse *Magel2* gene results in growth abnormalities similar to Prader-Willi syndrome. **Human Molecular Genetics**, v. 16, p. 2713-2719, 2007.

Bourdon R. M.; Brinks. J. S. Scrotal circumference in yearling Hereford bulls: adjustment factors, heritabilities and genetic, environmental and phenotypic relationships with growth traits. **Journal of Animal Science**, v. 62, p. 958-967, 1986.

Brown, P. R.; Miki, K.; Harper, D.B.; Eddy, E.M. A-kinase anchoring protein 4 binding proteins in the fibrous sheath of the sperm flagellum. **Biology of Reproduction**, v. 68, p. 2241-2248, 2003.

Chen, W.M.; Abecasis, G. R. Family-based association tests for genomewide association scans. **American Journal of Human Genetics**, v. 81, p. 913-926, 2007.

Dias, L. T.; Faro, L. E.; Albuquerque, L. G. Estimativas de herdabilidade para perímetro escrotal de animais da raça nelore. **Revista Brasileira de Zootecnia**, v. 32, p. 1878-1882, 2003.

Evans, J. L.; Golden, B. L.; Bourdon, R. M.; Long, K. L. Additive genetic relationships between heifer pregnancy and scrotal circumference in Hereford cattle. **Journal of Animal Science**, v. 77, p. 2621-2628, 1999.

Fortes, M. R. S.; Kemper, K.; Sasazaki, S.; Reverter, A.; Pryce, J. E.; Barendse, W.; Bunch, R.; McCulloch, R.; Harrison, B.; Bolormaa, S.; Zhang, Y. D.; Hawken, R. J.; Goddard, M. E.; Lehnert, S. A. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. **Animal Genetics**, in press, 2013.

Fortes, M. R. S.; Lehnert, S. A.; Bolormaa, S.; Reich, C.; Fordyce, G.; Corbet, N. J.; Whan, V.; Hawken, R. J.; Reverter, A. Finding genes for economically important traits: Brahman cattle puberty. **Animal Production Science**, v. 52, p. 143-150, 2012a.

Fortes, M. R. S.; Reverter, A.; Hawken, R. J.; Bolormaa, S.; Lehnert, S. A. Candidate genes associated with testicular development, sperm quality, and hormone levels of

inhibin, luteinizing hormone, and insulin-like growth factor 1 in Brahman bulls. **Biology of Reproduction**, 87:58, 2012b.

French, C. A.; Miyoshi, I.; Kubonishi, I.; Grier, H. E.; Perez-Atayde, A. R.; Fletcher, J. A. BRD4-NUT fusion oncogene: a novel mechanism in aggressive carcinoma. **Cancer Research**, v. 63 ,p. 304-307, 2003.

Garrick, D. J.; Taylor, J. F.; Fernando, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, 41:55, 2009.

Göllner, H.; Bouwman, P.; Mangold, M.; Karis, A.; Braun, H.; Rohner, I.; Del, R. e. y.; Besedovsky, H. O.; Meinhardt, A.; van den Broek, M.; Cutforth, T.; Grosveld, F.; Philipsen, S.; Suske, G. Complex phenotype of mice homozygous for a null mutation in the Sp4 transcription factor gene. **Genes Cells**, v. 6, p. 689-697, 2001.

Hagen, G.; Müller, S.; Beato, M.; Suske, G. Cloning by recognition site screening of two novel GT box binding proteins: a family of Sp1 related genes. **Nucleic Acids Research**, v. 20, p. 5519-5525, 1992.

Hu, Z. L.; Park, C. A.; Wu, X. L.; Reecy, J. M. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. **Nucleic Acids Research**, D871-9, 2013.

Huggins, C.; Moulton, S. H. Esterases of testis and other tissues. **Journal of Experimental Medicine**, v. 88, p. 169-179, 1948.

Karim, L.; Takeda, H.; Lin, L.; Druet, T.; Arias, J. A.; Baurain, D.; Cambisano, N.; Davis, S. R.; Farnir, F.; Grisart, B.; Harris, B. L.; Keehan, M. D.; Littlejohn, M. D.; Spelman, R. J.; Georges, M.; Coppieters, W. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. **Nature Genetics**, v. 43, 405-413, 2011.

Kinsella, R. J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; Kersey, P.; Flicek, P. Ensembl

BioMarts: a hub for data retrieval across taxonomic space. **Database (Oxford),** Bar030, 2011.

Kozlov, S. V.; Bogenpohl, J. W.; Howell, M. P.; Wevrick, R.; Panda, S.; Hogenesch, J. B.; Muglia, L. J.; Van Gelder, R. N.; Herzog, E. D.; Stewart, C. L. The imprinted gene *Magel2* regulates normal circadian output. **Nature Genetics**, v. 39, p. 1266-1272, 2007.

Langer, C.; Gansz, B.; Goepfert, C.; Engel, T.; Uehara, Y.; von Dehn, G.; Jansen, H.; Assmann, G.; von Eckardstein, A. Testosterone up-regulates scavenger receptor BI and stimulates cholesterol efflux from macrophages. **Biochemical and Biophysical Research Communications**, v. 296, p. 1051-1057, 2002.

Lango Allen, H.; Estrada, K.; Lettre, G.; Berndt, S. I.; Weedon, M. N.; Rivadeneira, F.; Willer, C. J.; Jackson, A. U.; Vedantam, S.; Raychaudhuri, S.; Ferreira, T.; Wood, A. R.; Weyant, R. J.; Segrè, A. V.; Speliotes, E. K.; Wheeler, E.; Soranzo, N.; Park, J. H.; Yang, J.; Gudbjartsson, D.; Heard-Costa, N. L.; Randall, J. C.; Qi, L.; Vernon Smith, A.; Mägi, R.; Pastinen, T.; Liang, L.; Heid, I. M.; Luan, J.; Thorleifsson, G.; Winkler, T. W.; Goddard, M. E.; Sin Lo, K.; Palmer, C.; Workalemahu, T.; Aulchenko, Y. S.; Johansson, A.; Zillikens, M. C.; Feitosa, M. F.; Esko, T.; Johnson, T.; Ketkar, S.; Kraft, P.; Mangino, M.; Prokopenko, I.; Absher, D.; Albrecht, E.; Ernst, F.; Glazer, N. L.; Hayward, C.; Hottenga, J. J.; Jacobs, K. B.; Knowles, J. W.; Kutalik, Z.; Monda, K. L.; Polasek, O.; Preuss, M.; Rayner, N. W.; Robertson, N. R.; Steinthorsdottir, V.; Tyrer, J. P.; Voight, B. F.; Wiklund, F.; Xu, J.; Zhao, J. H.; Nyholt, D. R.; Pellikka, N.; Perola, M.; Perry, J. R.; Surakka, I.; Tammesoo, M. L.; Altmaier, E. L.; Amin, N.; Aspelund, T.; Bhangale, T.; Boucher, G.; Chasman, D. I.; Chen, C.; Coin, L.; Cooper, M. N.; Dixon, A. L.; Gibson, Q.; Grundberg, E.; Hao, K.; Juhani Junttila, M.; Kaplan, L. M.; Kettunen, J.; König, I. R.; Kwan, T.; Lawrence, R. W.; Levinson, D. F.; Lorentzon, M.; McKnight, B.; Morris, A. P.; Müller, M.; Suh Ngwa, J.; Purcell, S.; Rafelt, S.; Salem, R. M.; Salvi, E.; Sanna, S.; Shi, J.; Sovio, U.; Thompson, J. R.; Turchin, M. C.; Vandenput, L.; Verlaan, D. J.; Vitart, V.; White, C. C.; Ziegler, A.; Almgren, P.; Balmforth, A. J.; Campbell, H.; Citterio, L.; De Grandi, A.; Dominiczak, A.; Duan, J.; Elliott, P.; Elosua, R.; Eriksson, J. G.; Freimer, N. B.; Geus, E. J.; Glorioso, N.; Haiqing, S.; Hartikainen, A. L.; Havulinna, A. S.; Hicks, A. A.; Hui, J.; Igl, W.; Illig, T.; Jula, A.; Kajantie, E.; Kilpeläinen, T. O.; Koiranen, M.; Kolcic, I.; Koskinen, S.; Kovacs, P.; Laitinen, J.; Liu, J.; Lokki, M. L.; Marusic, A.; Maschio, A.; Meitinger, T.; Mulas, A.; Paré, G.; Parker, A. N.; Peden, J. F.; Petersmann, A.; Pichler, I.; Pietiläinen, K. H.; Pouta, A.; Ridderstråle, M.; Rotter, J. I.; Sambrook, J. G.; Sanders, A. R.; Schmidt, C. O.; Sinisalo, J.; Smit, J. H.; Stringham, H. M.; Bragi Walters, G.; Widen, E.; Wild, S. H.; Willemsen, G.; Zagato, L.; Zgaga, L.; Zitting, P.;

Alavere, H.; Farrall, M.; McArdle, W. L.; Nelis, M.; Peters, M. J.; Ripatti, S.; van Meurs, J.B.; Aben, K. K.; Ardlie, K. G.; Beckmann, J. S.; Beilby, J. P.; Bergman, R. N.; Bergmann, S.; Collins, F. S.; Cusi, D.; den Heijer, M.; Eiriksdottir, G.; Gejman, P. V.; Hall, A. S.; Hamsten, A.; Huikuri, H. V.; Iribarren, C.; Kähönen, M.; Kaprio, J.; Kathiresan, S.; Kiemeney, L.; Kocher, T.; Launer, L. J.; Lehtimäki, T.; Melander, O.; Mosley, T. H.; Musk, A. W.; Nieminen, M. S.; O'Donnell, C. J.; Ohlsson, C.; Oostra, B.; Palmer, L. J.; Raitakari, O.; Ridker, P. M.; Rioux, J. D.; Rissanen, A.; Rivolta, C.; Schunkert, H.; Shuldiner, A. R.; Siscovick, D. S.; Stumvoll, M.; Tönjes, A.; Tuomilehto, J.; van, O. m. m. e. n.; Viikari, J.; Heath, A. C.; Martin, N. G.; Montgomery, G. W.; Province, M. A.; Kayser, M.; Arnold, A. M.; Atwood, L. D.; Boerwinkle, E.; Chanock, S. J.; Deloukas, P.; Gieger, C.; Grönberg, H.; Hall, P.; Hattersley, A. T.; Hengstenberg, C.; Hoffman, W.; Lathrop, G. M.; Salomaa, V.; Schreiber, S.; Uda, M.; Waterworth, D.; Wright, A. F.; Assimes, T. L.; Barroso, I.; Hofman, A.; Mohlke, K. L.; Boomsma, D. I.; Caulfield, M. J.; Cupples, L. A.; Erdmann, J.; Fox, C. S.; Gudnason, V.; Gyllensten, U.; Harris, T. B.; Hayes, R. B.; Jarvelin, M. R.; Mooser, V.; Munroe, P. B.; Ouwehand, W. H.; Penninx, B. W.; Pramstaller, P. P.; Quertermous, T.; Rudan, I.; Samani, N. J.; Spector, T. D.; Völzke, H.; Watkins, H.; Wilson, J. F.; Groop, L. C.; Haritunians, T.; Hu, F. B.; Kaplan, R. C.; Metspalu, A.; North, K. E.; Schlessinger, D.; Wareham, N. J.; Hunter, D. J.; O'Connell, J. R.; Strachan, D. P.; Wichmann, H. E.; Borecki, I. B.; van, D. u. i. j. n.; Schadt, E. E.; Thorsteinsdottir, U.; Peltonen, L.; Uitterlinden, A. G.; Visscher, P. M.; Chatterjee, N.; Loos, R. J.; Boehnke, M.; McCarthy, M. I.; Ingelsson, E.; Lindgren, C. M.; Abecasis, G. R.; Stefansson, K.; Frayling, T. M.; Hirschhorn, J. N. Hundreds of variants clustered in genomic loci and biological pathways affect human height. **Nature**, v. 467, p. 832-838, 2010.

Martin, L. C.; Brinks, J. S.; Bourdon, R. M.; Cundiff, L. V. Genetic effects on beef heifer puberty and subsequent reproduction. **Journal of Animal Science**, v. 70, p. 4006-4017, 1992.

McClure, M. C.; Morsci, N. S.; Schnabel, R. D.; Kim, J. W.; Yao, P.; Rolf, M. M.; McKay, S. D.; Gregg, S. J.; Chapple, R. H.; Northcutt, S. L.; Taylor, J. F. A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. **Animal Genetics**, v. 41, p. 597-607, 2010.

Pérez-Losada, J.; Sánchez-Martín, M.; Rodríguez-García, A.; Sánchez, M. L.; Orfao, A.; Flores, T.; Sánchez-García, I. Zinc-finger transcription factor Slug contributes to the function of the stem cell factor c-kit signaling pathway. **Blood**, v. 100, p. 1274-1286, 2002.

Pryce, J.E.; Hayes, B.J.; Bolormaa, S.; Goddard, M.E. Polymorphic regions affecting human height also control stature in cattle. **Genetics**, v. 187, p. 981-984, 2011.

Quinlan, A. R.; Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, p. 841-842, 2010.

R Development Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing 2008, Vienna, Austria**. Available at: <http://www.R-project.org>, Accessed on 14 oct. 2013.

Ricke, R. M.; Jeganathan, K. B.; Malureanu, L.; Harrison, A. M.; van Deursen J. M. Bub1 kinase activity drives error correction and mitotic checkpoint control but not tumor suppression. **Journal of Cell Biology**, v. 199, p. 931-949, 2012.

Rubin, C. J.; Zody, M. C.; Eriksson, J.; Meadows, J. R.; Sherwood, E.; Webster, M. T.; Jiang, L.; Ingman, M.; Sharpe, T.; Ka, S.; Hallböök, F.; Besnier, F.; Carlborg, O.; Bed'hom, B.; Tixier-Boichard, M.; Jensen, P.; Siegel, P.; Lindblad-Toh, K.; Andersson, L. Whole-genome resequencing reveals loci under selection during chicken domestication. **Nature**, v. 464, p. 587-591, 2010.

Sabatti, C.; Service, S. K.; Hartikainen, A. L.; Pouta, A.; Ripatti, S.; Brodsky, J.; Jones, C. G.; Zaitlen, N. A.; Varilo, T.; Kaakinen, M.; Sovio, U.; Ruokonen, A.; Laitinen, J.; Jakkula, E.; Coin, L.; Hoggart, C.; Collins, A.; Turunen, H.; Gabriel, S.; Elliot, P.; McCarthy, M. I.; Daly, M. J.; Järvelin, M. R.; Freimer, N. B.; Peltonen, L. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. **Nature Genetics**, v. 41, p. 35-46, 2009.

Satoh, T.; Hosokawa, M. The mammalian carboxylesterases: from molecules to functions**. Annual Review of Pharmacology and Toxicology**, v. 38, p. 257-288, 1998.

Stefan, M.; Portis, T.; Longnecker, R.; Nicholls, R.D. A nonimprinted Prader-Willi Syndrome (PWS)-region gene regulates a different chromosomal domain in trans but the imprinted pws loci do not alter genome-wide mRNA levels. **Genomics**, v. 85, p. 630-640, 2005.

Supp, D. M.; Witte, D. P.; Branford, W. W.; Smith, E. P.; Potter, S. S. Sp4, a member of the Sp1-family of zinc finger transcription factors, is required for normal murine growth, viability, and male fertility. **Developmental Biology**, v. 176, p. 284-299, 1996.

Toelle, V. D.; Robison, O. W. Estimates of genetic correlations between testicular measurements and female reproductive traits in cattle. **Journal of Animal Science**, v. 60, p. 89-100, 1985.

Utsunomiya, Y. T.; Carmo, A. S.; Carvalheiro, R.; Neves, H. H.; Matos, M. C.; Zavarez, L. B.; Pérez O'Brien, A. M. ; Sölkner, J.; McEwan, J. C.; Cole, J. B.; Van Tassell, C. P.; Schenkel, F. S.; da Silva, M. V. G. B.; Porto-Neto, L. R.; Sonstegard, T. S.; Garcia, J. F. Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. **BMC Genetics**, 14:52, 2013.

Van Melis, M. H.; Eler, J. P.; Rosa, G. J.; Ferraz, J. B.; Figueiredo, L. G.; Mattos, E. C.; Oliveira, H. N. Additive genetic relationships between scrotal circumference, heifer pregnancy, and stayability in Nellore cattle. **Journal of Animal Science**, v. 88, p. 3809-3813, 2010.

Vezzosi, D.; Bertherat, J. Phosphodiesterases in endocrine physiology and disease. **European Journal of Endocrinology**, v. 165, p. 177-188, 2011.

Yan, B.; Yang, D.; Brady, M.; Parkinson, A. Rat testicular carboxylesterase: cloning, cellular localization, and relationship to liver hydrolase A. **Archives of Biochemistry and Biophysics**, v. 316, p. 899-908, 1995.

Zimin, A. V.; Delcher, A. L.; Florea, L.; Kelley, D. R.; Schatz, M. C.; Puiu, D.; Hanrahan, F.; Pertea, G.; Van Tassell, C. P..; Sonstegard, T. S.; Marçais, G.; Roberts, M.; Subramanian, P.; Yorke, J. A.; Salzberg, S. L. A whole-genome assembly of the domestic cow, *Bos taurus*. **Genome Biology**, 10:R42, 2009.

CHAPTER 4 - Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods

Utsunomiya, Y. T.; Pérez O'Brien, A. M.; Sonstegard, T. S.; Van Tassell, C. P.; Carmo, A. S.; Mészarós, G.; Sölkner, J.; Garcia J. F.

PLoS ONE 8:e64280, 2013
DOI: 10.1371/journal.pone.0064280

1. Abstract

As the methodologies available for the detection of positive selection from genomic data vary in terms of assumptions and execution, weak correlations are expected among them. However, if there is any given signal that is consistently supported across different methodologies, it is strong evidence that the locus has been under past selection. In this paper, a straightforward frequentist approach based on the Stouffer Method to combine *P*-values across different tests for evidence of recent positive selection in common variations, as well as strategies for extracting biological information from the detected signals, were described and applied to high density single nucleotide polymorphism (SNP) data generated from dairy and beef cattle (taurine and indicine). The ancestral *Bovinae* allele state of over 440,000 SNPs is also reported. Using this combination of methods, highly significant ($P < 3.17 \times 10^{-7}$) population-specific sweeps pointing out to candidate genes and pathways that may be involved in beef and dairy production were identified. The most significant signal was found in the Cornichon homolog 3 gene (*CNIH3*) in Brown Swiss ($P = 3.82 \times 10^{-12}$), and may be involved in the regulation of pre-ovulatory luteinizing hormone surge. Other putative pathways under selection are the glucolysis/gluconeogenesis, transcription machinery and chemokine/cytokine activity in Angus; calpain-calpastatin system and ribosome biogenesis in Brown Swiss; and gangliosides deposition in milk fat globules in Gyr. The composite method, combined with the strategies applied to retrieve functional information, may be a useful tool for surveying genome-wide selective sweeps and providing insights in to the source of selection.

Key-words: Positive selection, Genome-wide scan, Meta-analysis, SNP, Cattle

2. Introduction

Selection changes the frequency of advantageous variants and their neighbor polymorphic sites, sweeping the genome and leaving patterns that become prevalent in a population despite chromosome recombination (SABETI et al., 2002). These patterns are broadly referred as signatures (or footprints) of selection, and many methods have been developed for identifying them from genomic data (OLEKSYK et al., 2010). The application of such approaches to dairy and beef cattle can help detecting chromosome regions that underwent not only natural but also anthropogenic selection, and that may be associated with traits of economic interest.

The available portfolio of methodologies varies in terms of the underlying selection processes assumed, the age of the sweep, and if the test is performed within-population or depends on population comparisons (Table 1). In this scenario, one may expect that correlations among different tests are weak. However, if there is any given signal consistently supported across different methodologies, it may be strong evidence that the locus has been under past selection.

Recently, Grossman et al. (2010) stated that ‘‘If each signature provides distinct information about selective sweeps, combining the signals should have greater power for localizing the source of selection than any single test’’. Driven by this thought, they developed a Bayesian method for combining *P*-values from different approaches, namely Composite of Multiple Signals (CMS), which was capable to discriminate causal variants from neutral markers in simulated data. Application of CMS to real data led to the discovery of evidence of recent positive selection in *LARGE* and *IL2* in Nigeria human population, genes that were previously incriminated in resistance to Lassa Fever (ANDERSEN et al., 2012).

Although suitable for analysis of human populations, CMS is still challenging to be applied to cattle genomic data, as the computation of likelihood tables requires coalescent simulations using calibrated demographic models in an attempt to mimic the empirical data. Despite availability of good models for cattle history (MURRAY et al., 2010), uncertainties around the model and specific recent events that happened during breed formation makes difficult matching the simulations to the real data.

Table 1. Types of signatures of selection detectable from genomic data. Ages of selection are based on estimations for human data in years, assuming a generation interval of 25 years (OLEKSYK et al., 2010).

Type of signature	Detectable pattern	Methodologies	Underlying selection phenomena	Population level	Age of selection (generations)	References
Function-altering mutation	Changes in non-synonymous to synonymous variation ratio in the open reading frame of a coding region	$\omega = D_n/D_s$	positive and purifying selection	Within species	> 40,000	Nielsen et al. (1988)
Local genetic diversity depression	Deficit of local heterozygosity compared to the rest of the genome	*ZHp*, SNP heterozygosity	positive selection	Within populations	< 10,000	Rubin et al. (2010); Oleksyk et al. (2010)
Change in the allele frequency spectrum	Increase in the frequency of derived alleles	*ΔDAF*, *Tajima's D*, *Fu and Li's D-test*, *Fay and Wu's H-test*, *CLR*	positive selection	Within and between populations	< 3,200	Grossman et al. (2010); Tajima et al. (1989); Fu and Li (1993); Fay and Wu (2000); Williamson et al. (2007)
Population differentiation	Difference in the allele frequencies between populations	F_{ST}	positive and balancing selection	Between populations	< 3,000	Weir and Cockerham (1984)
Extended haplotype homozygosity	LD persistency and unusual long-range haplotypes	*LRH*, *iHS*, *XP-EHH*, *Rsb, ΔiHH*, *varLD*	positive selection	Within and between populations	< 1,200	Sabeti et al. (2002); Voight et al. (2009); Sabeti et al. (2007); Tang et al. (2007); Grossman et al. (2010); Ong and Teo (2010)

This paper describes and applies to dairy and beef cattle data a straightforward frequentist meta-analysis approach for combining *P*-values across different tests for footprints of recent positive selection in genome-wide single nucleotide polymorphism (SNP) data, targeting common, moderate frequency variants. Two between and two within population tests for selection sweeps are covered, divided into three different categories: extended haplotype homozygosity (*EHH*), change in the allele frequency spectrum, and local heterozygosity depression. Strategies for assigning relevant SNPs to genes are also described, allowing for exploration of the biological meaning of the findings and facilitating hypothesis generation. Additionally, the ancestral *Bovinae* allele state of over 440,000 SNPs is reported.

3. Material and methods

3.1. Samples and quality control

Genotypes for Illumina® BovineHD Genotyping BeadChip assay of Angus (ANG), Brown Swiss (BSW), Gyr (GYR) and Nellore (NEL) individuals were available for prospection of selection sweeps. Details on sample size and data source for each breed can be found in Table 2. Only autosome markers (n = 742,910) were included into the analyses. Markers were removed from the dataset if they did not exhibit: 1) minor allele frequency (MAF) greater than or equal to 0.03, 2) *P*-value for Hardy-Weinberg Equilibrium (HWE) greater than or equal to $1x10^{-6}$ or 3) Call rate (CR_{SNP}) greater than or equal to 90%. After the SNP quality control (QC), individuals exhibiting call rate (CR_{IND}) below 90% were also removed. This procedure was performed for each breed genotype's dataset in parallel using *PLINK* (PURCELL et al., 2007). In order to mitigate relatedness in the dataset, individuals were further investigated for the proportion of alleles shared identically by descent (IBD) using *PLINK*. Potential parent-offspring, half-siblings and duplicate pairs were conservatively removed (see Appendix B for details). Markers commonly passing QC in all four breeds were then overlapped. As the final SNP set consisted of markers passing QC with relatively small amount of missing data, and most of the methods for

the detection of selection sweeps do not accommodate missing values, an imputation procedure was adopted to fill the existing missing genotypes. For this purpose, *fastPHASE* software was used (SCHEET et al., 2006) with the following arguments: *-H-4 -K10 -T10 -C25*.

3.2. Ancestral allele discovery

Since some methodologies for detecting positive selection rely on the comparison of the recombination breakdown between haplotypes carrying the ancestral and the derived allele (SABETI et al., 2002; VOIGHT et al., 2009; SABETI et al., 2007), ancestral allele states were assessed using outgroup species assumed to be derived from a common founder *Bovinae* species that included 2 Gaur (*Bos gaurus*), 6 Water Buffalo (*Bubalus bubalis*) and 2 Yak (*Bos grunniens*) with genotypes derived from the same assay. Genotypes for the three outgroup *Bovinae* species were pooled into a single dataset. Markers with a CR_{SNP} of 100% (i.e., the SNP probe designed to hybridize bovine DNA also recognizes other *Bovinae* species, meaning that the target sequence is within a syntenic block across the outgroups and may have been inherited from a common ancestor) and MAF = 0 (i.e., monomorphic markers, being the one single allele present likely to be the common ancestral variant) were sought. For each case, the major allele (frequency = 100%) was determined as ancestral. The final SNP set was then defined and included markers passing QC with ancestral allele information available.

3.3. Genome-wide scan methods for positive selection

3.3.1. Long-range haplotype based methods

The two methodologies described here are based on the concept of Extended Haplotype Homozygosity (*EHH*) (SABETI et al., 2002), and were applied using the *rehh* package in *R* (GAUTIER & VITALIS, 2012) with minor adaptations to the source code. As the basis for the two tests, the integrated *EHH* for the ancestral allele (iHH_A), derived allele (iHH_D) and SNP site (*iES*) was calculated for each marker. *EHH*

Method 1: Voight et al. (2006) described a within population score for the ratio between iHH_A and iHH_D, called Integrated Haplotype Score (*iHS*):

$$iHS = ln\left(\frac{iHH_A}{iHH_D}\right)$$

As *iHS* distribution is approximately normal, the scores are divided into 20 equally sized bins according to their derived allele frequencies, and then standardized to have mean 0 and variance 1. The scores reflect how unusual the haplotypes containing the ancestral (positive values) and derived (negative values) allele are, relative to the entire genome. As both tails from the distribution were of interest, two-sided *P*-values were derived as $1 - 2|\Phi(iHS) - 0.5|$ from the Gaussian cumulative density function. *EHH* Method 2: Tang et al. (2007) defined *Rsb*, a between populations test, as:

$$Rsb = ln\left(\frac{iES_{pop1}}{iES_{pop2}}\right)$$

The outcome also resembles a normal distribution. Unlike *iHS*, the standardization procedure recommended by Tang et al. (2007) does not divide scores into bins and uses the median instead of the mean. Positive values suggest selection in the population used in the numerator, while negative values indicate signals in the population used as denominator. For each pair of breeds, *Rsb* scores were calculated using the standardization procedure recommended by Tang et al. (2007). As every population was used both as numerator and denominator, one-sided upper tail *P*-values were derived from the normal cumulative density function.

3.3.2. Change in the allele frequency spectrum

Grossman et al. (2010) described a simple method based on the difference in the derived allele frequency between populations (*ΔDAF*). Values range from -1 to 1

and are normally distributed. *ΔDAF* scores were standardized using the distribution's mean and standard deviation, and one-sided upper tail *P*-values were obtained.

3.3.3. Local heterozygosity depression based method

Rubin et al. (2010) defined and applied a Z-score test for local heterozygosity depression (*ZHp*) on whole genome sequence data of domestic chicken, which basically expresses how much the expected heterozygosity in chromosome windows deviate from the average genome heterozygosity. The approach was adapted to each SNP site and computed using the observed instead of the expected heterozygosity values. The values were standardized to produce mean 0 and variance 1. For this method, negative values were of interest and the resulting site heterozygosity scores were multiplied by -1 in order to switch their direction, yielding a new statistic called *SHp* (i.e. site *ZHp*). One-sided upper tail *P*-values were obtained for each score.

3.3.4. Meta-analysis of multiple tests

As all applied methodologies had *P*-values retrieved from normal distributions with same parameters (mean 0 and variance 1), the weighted version of Stouffer method was adapted for the combination of *Z*-transformed *P*-values, as reviewed by Whitlock (2005). For each marker and each test i, the respective *P*-value was transformed into a *Z*-score by $Z_i = \Phi^{-1}(1 - P_i)$. Within population tests were performed only once per breed, hence their respective weight ω_i was set to 1. For each comparison of between population tests, the Z-score was weighted to $1/n$, where n is the number of comparisons. Then, the combined statistic of k tests, for each SNP in each breed, was defined as:

$$meta - SS = \frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt{\sum_{i=1}^{k} \omega_i^2}}$$

The *meta-SS* (stands for Meta-analysis of Selection Signals) scores were referred back to the standard normal distribution in order to obtain combined significance values, which were intended to address either the combination of information among different, independent tests can reject the shared null hypothesis (neutral marker). Significance level for genome-wide *meta-SS P*-values was based on a Bonferroni threshold ($\alpha = 0.05/n_{SNP}$).

3.4. Functional annotation

For every peak crossing the significance line, three different strategies for the annotation of functional features were applied, based on the genomic coordinates from the *UMD3.1* assembly (ZIMIN et al., 2009). Strategy 1: Since any given gene harboring signals is a direct candidate, the presence of significant intragenic SNPs was checked in the *Ensembl Variation 67* database using the *BioMart tool* (KINSELLA et al., 2011). Strategy 2: The closest gene in the vicinity of the most relevant SNP of a given peak could be responsible for the signal. Hence, the most significant SNP from each observed peak was isolated and the closest gene to it was mapped using the *ClosestBed* algorithm from the *BedTools* software (QUINLAN & HALL, 2010). Strategy 3: Since there are cases where variants from multiple genes in linkage disequilibrium (LD) with the marker contribute to the signal together, due to the fact that functionally related genes are often spatially close to each other (TANG et al., 2007), the third approach was based in a window scheme to capture genes that were potentially in LD with the significant SNP. The derived gene lists were processed in *DAVID* (HUANG et al., 2009a; HUANG et al., 2009b) for annotation of functional terms. Although *DAVID* provides means for enrichment analysis, with significance tests for overrepresented terms, the inclusion criterion of functional terms was solely based on existence of information. Finally, the *Enrichment Map Cytoscape plugin* (METICO et al., 2010) was used to build networks of inter-related terms based on the number of genes shared between terms, i.e., no hypothesis or significance test was applied, being the networks strictly descriptive. Terms were drawn as nodes (circles). Edges linking nodes represented gene sharing, and their

thickness, the degree of gene set overlap (i.e., proportional to the number of genes being shared). An extended description of this section is provided in Appendix B.

4. Results

4.1. Ancestral allele discovery

By assessing the outgroup species genotypes, an average CR_{IND} of 83.79%, 96.93%, 94.87% and 88.63% for Water Buffalo, Yak, Gaur and pooled data was observed, respectively. From the initial set of 742,910 autosomal markers, considering only markers perfectly typed across the pooled outgroup samples (CR_{SNP} = 100%), a total of 559,663 SNP probes were successfully hybridized (71.94%), and 111,376 SNPs were polymorphic (MAF > 0). Hence, a total of 448,287 SNPs (56.75%) had their ancestral allele determined, being provided as a TSV file (available at: http://dx.doi.org/10.1371/journal.pone. 0064280).

4.2. Quality control

Number of SNPs passing QC was 579,470, 554,826, 485,655 and 461,702 for ANG, BSW, GYR and NEL, respectively. Overlapping of the four SNP lists retrieved a final set of 281,994 markers, from which 157,702 had ancestral allele information available. Even with the drastic drop in the number of SNPs, the intermarker distance mean and median were 15.94 kb and 6.43 kb, respectively, superposing the median spacing of 37 kb declared for the BovineSNP50 assay (MATUKUMALLI et al., 2009). These findings indicated that the overall marker coverage was satisfactory, although generation of local gaps by QC was observed. No individuals were removed due to QC. The number of remaining samples for each breed, after duplicates and first degree relationship removal, was: 24 for ANG, 44 for BSW, 23 for GYR and 581 for NEL. As NEL exhibited a sample size much larger than the other breeds, 45 individuals were sampled from the total (Appendix B). Details on the final base dataset used for all further analyses can be found in Table 2.

Table 2. Description of cattle genotypes available for analysis before (BF) and after (AF) filtering for cryptic relatedness and quality control.

Breed	Code	Subspecies	Purpose	HapMap[a]		BOKU[b]		ZGC[c]		Total	
				BF	AF	BF	AF	BF	AF	BF	AF[f]
Angus	ANG	*Bos taurus*	Beef	27	24	0	0	0	0	27	24
Brown Swiss	BSW	*Bos taurus*	Dairy	24	13	48	31	0	0	72	44
Gyr	GYR	*Bos indicus*	Dairy	30	23	0	0	0	0	30	23
Nellore	NEL	*Bos indicus*	Beef	35	24	0	0	691	21[d]	726	45

[a]The Bovine HapMap Consortium (2009)
[b]University of Natural Resources and Life Sciences, Vienna.
[c]Zebu Genome Consortium.
[d] The actual number of NEL samples passing control criteria was 581: 557 for ZGC and 24 for HapMap. In order to avoid an unbalanced dataset, we decided to keep a final set of 45 NEL: all 24 HapMap samples plus 21 randomly chosen ZGC samples.
[f] Final base dataset used for the selective sweep analyses.

4.3. Identification of selection signals and functional annotation

All performed tests for footprints of selection resembled a normal distribution (Appendix C - Figure 1C) and genome-wide *Z*-transformed *P*-values were weakly correlated, satisfying the independence condition for meta-analysis (Appendix C - Figure 2C). Genome-wide distribution of *meta-SS P*-values and the closest genes to the top of the peaks can be found in Figure 1. The number of SNP with *P*-value crossing the genome-wide significance ($P < 3.17 \times 10^{-7}$) was: 153 for ANG, 212 for BSW, 3 for GYR and 13 for NEL. The most significant SNP was found in BSW ($P = 3.82 \times 10^{-12}$), and is an intronic variation in Cornichon homolog 3 gene (*CNIH3* - ENSBTAG00000044171), located at BTA16:28478192.

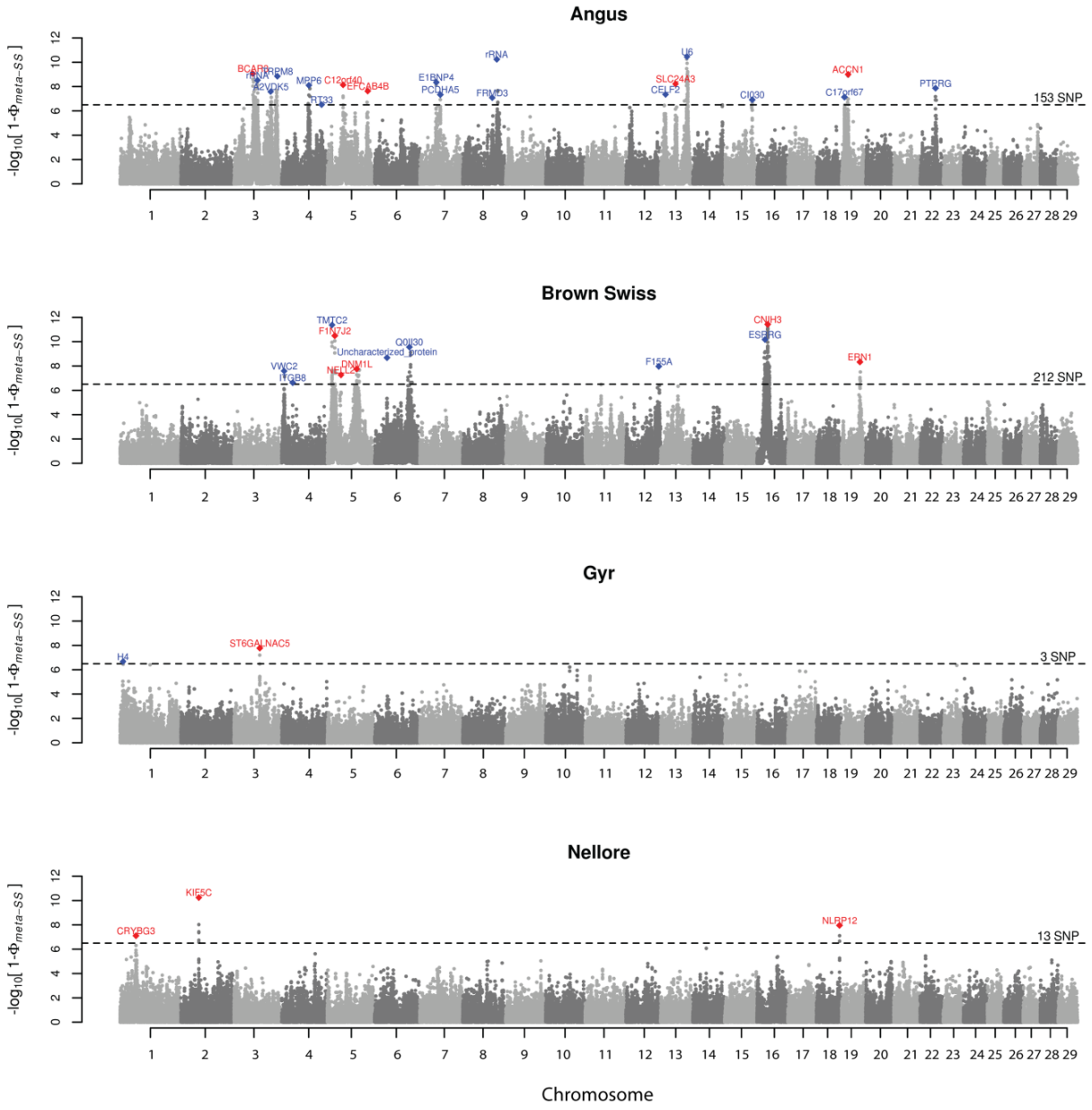

Figure 1. Manhattan plots of genome-wide meta-SS -$\log_{10}$(*P*-values) for Angus, Brown Swiss, Gyr and Nellore breeds. Number of SNPs indicated represents count of markers crossing the significance line ($P < 3.17 \times 10^{-7}$). Red and blue diamonds are intragenic and intergenic top SNPs on peaks, respectively.

In order to illustrate the potentiality of combining signals resulting from different methodologies for the detection of positive selection, a regional plot of *P*-values for each of the individual tests for the *CNIH3* region in BSW (candidate for being selected) and NEL (candidate for being neutral) was provided in Figure 2. For the same genomic region, two extra graphics were provided: 1) a *EHH* decay plot, showing the decrease of the probability of IBD as a function of the distance from the core SNP site (i.e., the *CNIH3* intronic SNP) for both the haplotypes containing the derived and ancestral alleles, and 2) a bifurcation diagram for the haplotypes containing the derived allele, representing the breakdown of LD at increasing distances from the core allele (in this case, the derived allele) at a given core SNP (in this case, the *CNIH3* intronic SNP). It can be seen from the BSW and NEL comparison that the signal of the unusual derived allele long haplotype in BSW, revealed by the *meta-SS* statistics, is not detectable in NEL. It is noticeable, by the shape of the SNP significances distribution in the *meta-SS* scatter plot, that *iHS* and *Rsb* had higher influence in the composite test, and the combination of methods penalized SNPs with little statistical support.

The number of genes directly harboring significant SNPs was 20 for ANG, 27 for BSW, 1 for GYR and 3 for NEL (the full list can be viewed at: http://dx.doi.org/10.1371/journal.pone.0064280). Two synonymous exonic SNP for ANG and BSW, one non-synonymous variation (BTA7:42652319, Ala->Thr) for a gene of the olfactory receptor family (*LOC524290*/*OR2W3* - ENSBTAG00000025293) in ANG (P = 7.65x10^{-9}) and a 3'UTR variation (BTA2:47315215) for the *KIF5C* (kinesin family member 5C - ENSBTAG00000018125) gene in NEL (P = 2.68x10^{-7}) were found. All other variants within genes were located in introns. The application of the LD-window approach (Strategy 3) retrieved SNP windows with an average size of 576.8 kb overall breeds, and the largest window spanned 1.83 Mb. Total number of genes within windows included in each breed specific list was: 309 for ANG, 177 for BSW, 4 for GYR and 14 for NEL (full lists can be found at: http://dx.doi.org/10.1371/journal.pone.0064280).

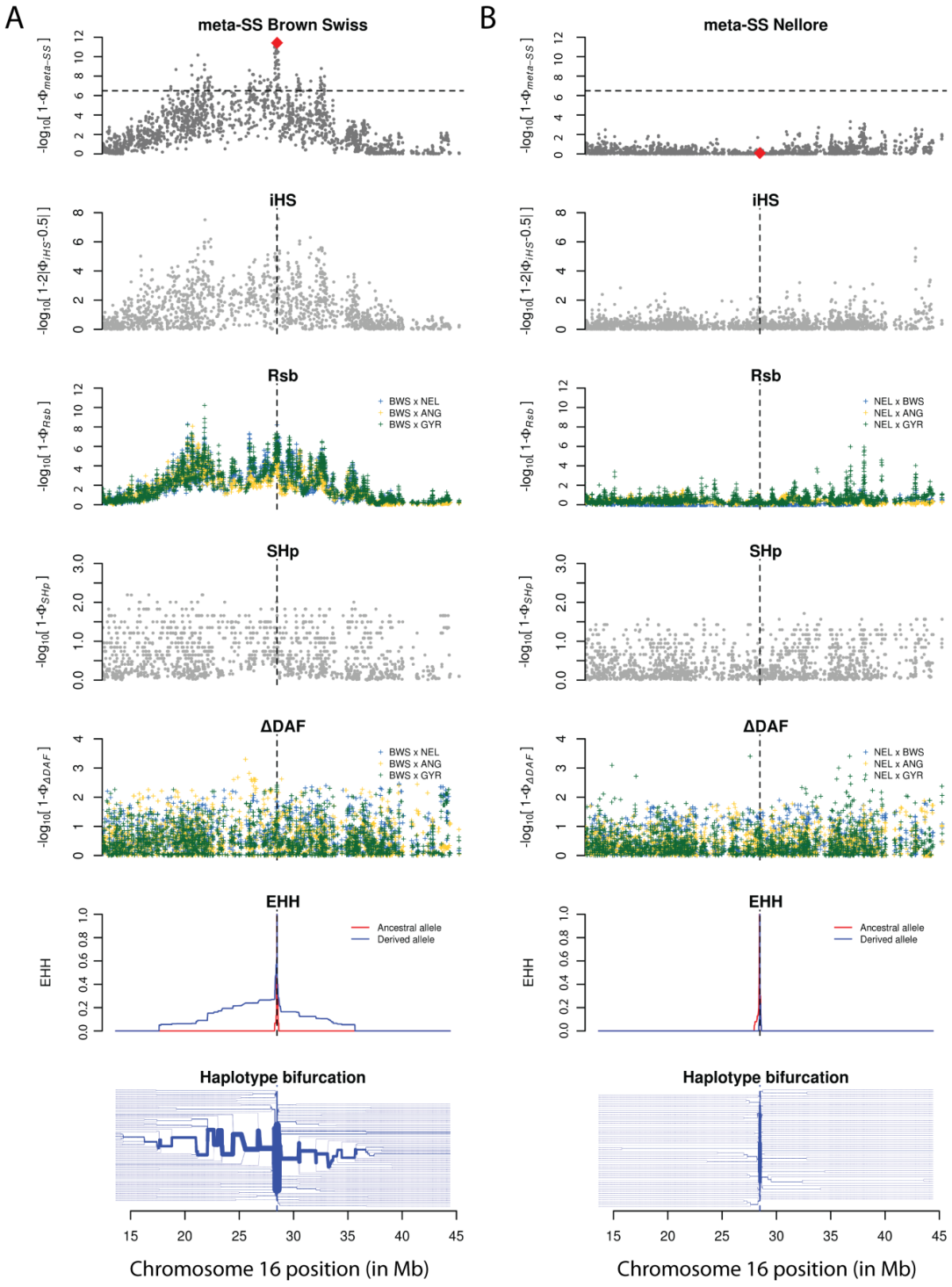

Figure 2. *meta-SS*, component tests, *EHH* and derived allele bifurcation for *CNIH3* in Brown Swiss (A) and Nellore (B). Vertical dashed lines and red diamonds represent the position of the intronic SNP detected as highly significant in Brown Swiss (BTA16:28478192, $P = 3.82 \times 10^{-12}$). Horizontal dashed lines mark the Bonferroni significance threshold ($P = 3.17 \times 10^{-7}$).

Networking of functional terms from ANG gene list (Figure 3A) revealed three groups: 1) immune response related genes, involved with chemokine and cytokine activity; 2) transcription activity, comprising the biosynthesis of ribonucleoproteins, transcription activation and aminoacylation of tRNA with L-histidine residuals; and 3) glucolysis and gluconeogenesis pathways. For BSW, a network related to post-transcriptional modifications of rRNAs (mostly methylation of adenosine residuals) and another involved with Calpain (Figure 3B) were observed. A significant intronic SNP (BTA16:27801014, $P = 2.61x10^{-7}$) was detected in the Calpain 2 (m-Calpain - ENSBTAG00000012778) catalytic subunit, which may be capturing the signal of a causal untyped variant under selection. Due to a low number of genes mapped, it was not possible to build a network of functional terms for GYR and NEL. Across all lists, a total of 69 genes (13.69%) had no functional term associated to them, being either uncharacterized proteins or novel RNAs with no functional record available. All DAVID annotation chart reports are provided at: http://dx.doi.org/10.1371/ journal.pone.0064280).

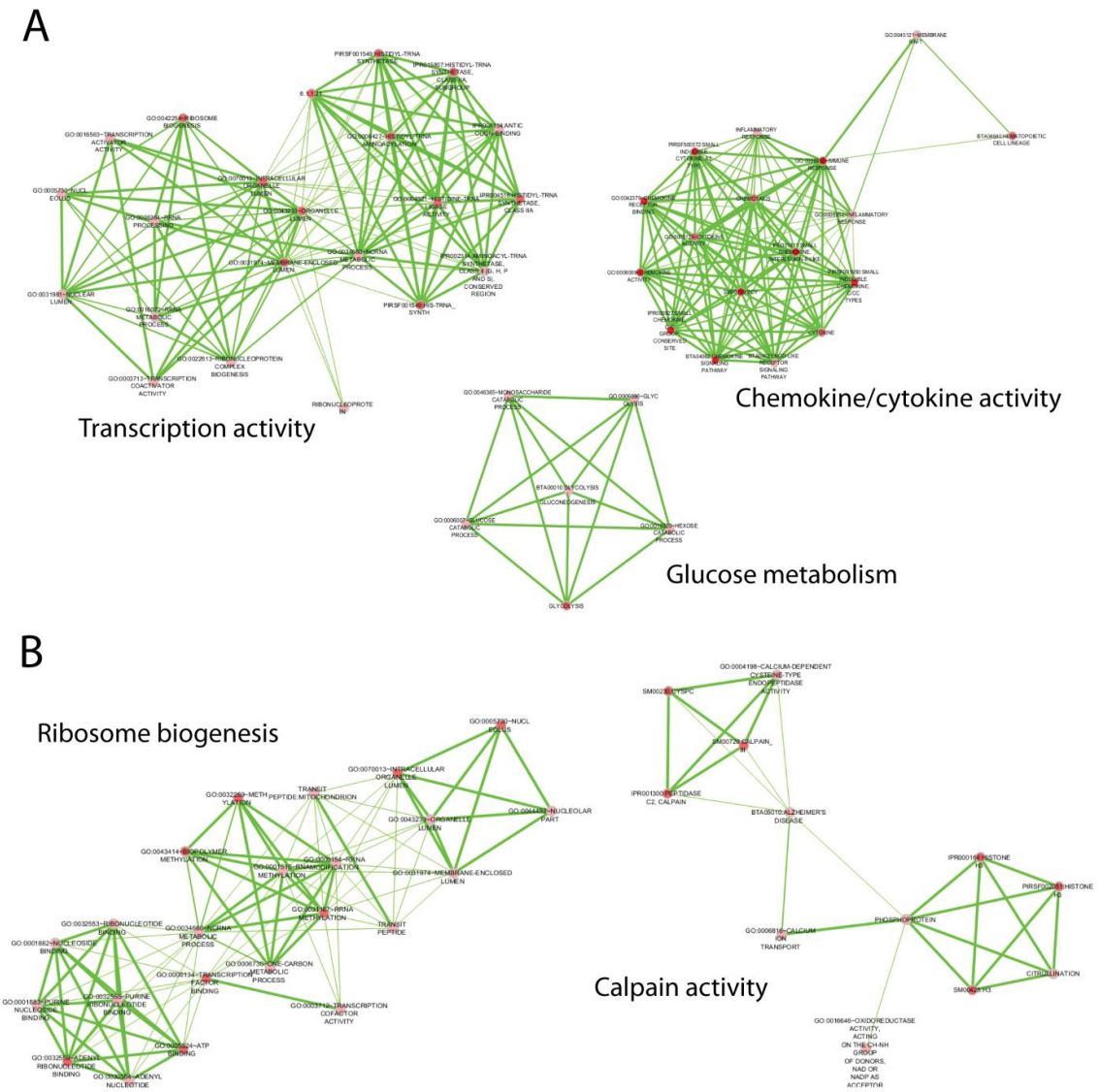

Figure 3. Descriptive Network of functional terms in Angus (A) and Brown Swiss (B). Nodes (red circles) are annotated functional terms. Edges connecting nodes represent gene share, being thickness proportional to the number of genes shared between terms (i.e., the degree of gene set overlap).

5. Discussion

Concordances among *EHH* based tests seemed to have led the composite statistics in most cases, and disagreements between *Rsb* and *iHS* scores showed severe drop in significance support. It was noticed that *SHp* and *ΔDAF* did not

contribute much towards spatial resolution individually, but they did help pinpointing SNP when blended with the other methods. This was in line with observations made by Grossman et al. (2010) when applying the ΔDAF method, which despite the little power to localize the sweep alone, showed to better distinguish selected from neutral variants in that study. All methods applied in this study were deemed capable of identifying recent sweeps, as well as signatures dating back up to a few thousand generations (OLEKSYK et al., 2010). Considering that the significance of the combined test was mainly influenced by *EHH* based tests, and that cattle generation interval vary between 3 and 5 years, the methodology applied could have identified sweeps that happened as far as 6,000 years ago (1,200 cattle generations). Although this comprises most of cattle domestication history, the majority of the signals detected are more likely to have arisen during breed formation, which goes up to some hundreds of years ago (AJMONE-MARSAN et al., 2010). This argument is based on two observations: 1) the meta-analysis method applied herein focused on breed-by-breed test integration, which may have favored the detection of breed-specific recent signatures; 2) for strong positive selective sweeps, which may have happened early in cattle domestication, the favored allele is expected to be nearly fixed across cattle breeds, and intrinsic factors of the present study contributed to the underrepresentation of fixed loci within the dataset used.

One factor that contributed to the underrepresentation of fixed loci in the dataset used is related to the SNP assay. As the SNP ascertainment strategies for the design of bovine arrays were focused on developing marker panels of common variations to support genome-wide association applications, and relied on sequence data of most major breeds for variation detection (MATUKUMALLI et al., 2009), the presence of SNP sites harboring rare variants (i.e. nearly fixed loci across cattle breeds) is scarce. Even if sequence data was used, the bovine reference genome available for detecting variants is the domesticated type (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al., 2009), and sites of variation that underwent strong positive selection during domestication are probably difficult to be identified, as the unselected variant may be very rare. For instance, Rubin et al. (2010) sought genome-wide heterozygosity depression in chicken using low coverage whole genome sequence data of DNA pools of domestic and wild lines,

and the reference genome of what is considered to be the ancestral type (*Gallus gallus*). The strategy allowed for the detection of genome regions that were nearly fixed in the domesticated lines and exhibited low identity to the wild-ancestor haplotypes, suggesting selection sweeps during domestication.

Another important factor contributing to the low representation of rare variants in the present study is the filtering of SNP with moderate allele frequencies in all breeds (MAF < 0.03), which may have made the detection of selection sweeps dating back to early cattle domestication unlikely. Nevertheless, the strategies adopted seemed to be capable of detecting footprints of recent positive selection, which may be anthropogenically ascertained by breeding and related to milk and meat production. The functional findings discussed later support this hypothesis.

The most significant signal found comes from the *CNIH3* gene in BSW ($P = 3.82x10^{-12}$). Figure 4 shows all present date known and predicted relationships of human *CNIH3* with other proteins assessed by data integration in the STRING 9.0 database (SZKLARCZYK et al., 2011), which indicates direct interaction of *CNIH3* with multiple amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) selective glutamate receptors (*GRIA1*, *GRIA2*, *GRIA3*, *GRIA4* and *GRIK1*). *CNIH3* regulates the trafficking and gating properties of AMPA receptors in the central nervous system (SHI et al., 2010), which were previously shown to participate in luteinizing hormone (LH) secretion (BRANN & MAHESH, 1997). Sugimoto et al. (2010) detected a single amino acid substitution (Ser -> Asn) in the bovine *GRIA1* that leads to decreased release of gonadotropin-releasing hormone (GnRH) and slower pre-ovulatory LH surge, making carrier cows less responsive to superovulation hormone treatment. Sugimoto et al. (2010) sequenced *GRIA1* in Japanese Black and Holstein commercial sires and found no departures from HWE in the locus, meaning that there is no evidence of selection pressure on the reported variants in either breeds. The signal on *CNIH3* found in the present study suggests that at least the underlying pathway has suffered recent selection pressure in BSW, although the selection force is unknown.

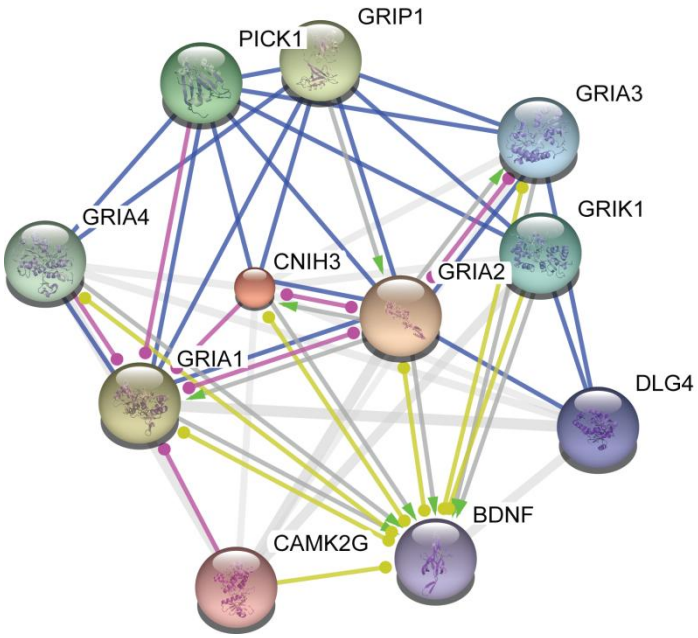

Figure 4. Protein network of human *CNIH3*, according to STRING 9.0 action view. Nodes are proteins; edges and arrows indicate interaction. Blue edges: binding; green arrows: activation; pink edges: post-translational modification; yellow edges: expression.

ANG exhibited three groups in the network map, one addressing chemokine/cytokine activity, a second with components of the transcription machinery and another related to glucolysis/gluconeogenesis. Both transcription activity and glucose metabolism are broad themes to be hypothesized, but it is possible that they have faced recent selection for high metabolic efficiency relative to increased meat yield and fat deposition. Regarding cytokines and chemokines, it has been found that they modulate different stages of muscle cell development (GADIENT & PATTERSON, 1999; ZOICO & ROUBENOFF, 2002). In a recent work, Zhao and collaborators (2012) found evidence that RNA expression of genes involved with acute inflammatory response has high influence in meat tenderness in Angus cattle. They observed that chemokines and cytokines genes, including chemokine C-C ligand 8 present in our gene list (ENSBTAG00000014113, BTA19), were deregulated in animals submitted to a surgical procedure, which in turn showed higher Warner-Bratzler shear force in beef samples after slaughter compared to the

control group, suggesting that they play important role in muscle metabolism, either in vivo or in postmortem proteolysis regulation. These findings support that genes participating in chemokine/cytokine activity are under selection in ANG cattle.

The calpain-calpastatin system is a proteolytic complex that has also been largely incriminated in postmortem meat tenderization in beef cattle (KOOHMARAIE & GEESINK, 2006). However, evidence of selection for components related to this system in the BSW data is somewhat surprising. Based on proteome analysis, Kuhla et al. (2011) proposed a model in which the muscle breakdown provides substrates for milk production in early lactation, being a key mechanism in the nutritional imbalance of high-yielding dairy cows. Although Kuhla et al. (2001) did not mention the calpain-calpastatin system, this may be one hypothesis for the overrepresentation of related terms found. Alternatively, Arnandis et al. (2012) has shown that calpains are responsible for mitochondrial and lysosomal membrane permeabilization during lysosomal-mediated mammary epithelial cell death in mice. Also, milk yield is known to decline as a function of many factors after peak lactation, including decrease in alveolar secretory epithelial cell number due to programmed cell death (WILDE et al., 1997). These evidences, together with the functional terms found, bring a second hypothesis that calpain-related genes are candidates under selection for lagged or mild post-peak lactation mammary gland involution in BSW. Both hypotheses point to the calpain-calpastatin system as a new target pathway involved in lactation dynamics in dairy cows.

Another intriguing candidate pathway pointed out by the present study is the ribosome biogenesis in BSW, more particularly the step involving methylation of rRNA. The addition of a methyl group to the 29-hydroxyl group of the backbone ribose is a conserved type of post-transcriptional RNA modification (KISS, 2002), and is an essential step in ribosome assembly. Most 29-O-methylated sites occur in functionally important regions of rRNAs and may influence ribosome structure and function (DECATUR & FOURNIER, 2002). It has been found that expression of ribosome components did not increase, and some of them had a slight decrease, during lactation in bovine mammary, which may be due to prioritization of synthesis of milk-specific mRNAs (BIONAZ & LOOR, 2011). Thus, since the anabolic demand in lactation is not accompanied by increase in the expression of ribosome

components, rRNA post-transcriptional modifications may play an important role in the translation efficiency of milk-specific proteins during lactation.

It was found an intronic signal for *ST6GALNAC5* (ENSBTAG00000007309) in GYR (P = 1.24x10^{-9}). *ST6GALNAC5* is involved in the synthesis of gangliosides, more particularly the GD1α in the brain (MOMOEDA et al., 2007). Gangliosides are glycosphingolipids containing one or more sialic acid residues in their structure, mainly n-acetylneuraminic acids. Some types of gangliosides can be found as components of the membrane fraction of the milk fat globule, which derives from the apical plasma membrane of secretory cells in the lactating mammary gland (BODE et al., 2004). Prolactindependent deposition of GD1α gangliosides in the milk fat globules of mice (comprising up to 80.5% of the total milk lipidbound sialic acid at the 3rd day of lactation) has been reported as a result of the expression of *ST6GALNAC5* during lactation, and may be an important source of GD1α for the developing neonate brain (MOMOEDA et al., 2007). These evidences raise the hypothesis that *ST6GALNAC5* has been indirectly selected in GYR via percentage of fat in the milk.

The present study found substantially fewer evidences of recent selection in GYR and NEL, relative to BSW and ANG. When only within population tests were combined in the meta-SS statistics, GYR and NEL exhibited considerable numbers of selective sweeps, but still less than the taurine breeds analyzed (Appendix C - Figure 3C). However, when only between populations tests were combined, the indicine breeds showed a severe drop in signals (Appendix C - Figure 4C). As tests based on LD persistency and unusual long-range haplotypes were an important part of the composite statistics, the decreased number of sweeps found in the indicine breeds could be explained by differences in haplotype block structure and extent of LD across taurine and indicine breeds. In fact, ANG and BSW were shown to have greater mean haplotype block size and average LD than GYR and NEL (VILA-ANGULO et al., 2004). Thus, the higher extended haplotype homozygosity of taurine breeds may have masked the detection of selective sweeps in the genomes of indicine breeds.

Many studies on signatures of selection in cattle have been published in recent years, and known genomic regions under selection are often used in literature

as ‘confirmatory’ loci in order to validate new findings. Examples of such loci are the melanocortin 1 receptor gene (*MC1R* - ENSBTAG00000023731), responsible for the black/red coat color in ANG (MATUKUMALLI et al., 2009; KUNGLAND et al., 1995; STELLA et al., 2010), and the Mast/stem cell growth factor receptor gene (*KIT* - ENSBTAG00000002699), incriminated in the ‘piebald’ spotted coat-color in Hereford and BSW (GROSZ & MACNEIL, 1999; STELLA et al., 2010). *MC1R* is located at BTA18: 14757332-14759082, and *KIT* is located at BTA6:71796318-71917431 in the *UMD v3.1* assembly. Both *KIT* and *MC1R* regions were underrepresented in SNP coverage in the present study due to QC effects and ancestral allele information availability, and gaps spanning BTA6 71.7–72.4 Mb and BTA18 14.0–15.0 Mb were observed. These observations could justify the absence of significant signals for *KIT* or *MC1R*. However, other studies searching for selective sweeps in these breeds also did not report signals in *KIT* and *MC1R* regions in BSW and ANG (THE BOVINE HAPMAP CONSORTIUM, 2009; QANBARI et al., 2011), respectively.

Some of the putative loci under selection detected herein were compared to previous studies, more particularly the signals found in BSW, since information on signatures of selection is more abundant in taurine dairy cattle. The topology of *iHS* - $\log_{10}$(*P*-values) across BTA 4, 5, 16 and 19 reported by Schwarzenbacher et al. (2012) was noticeably similar to the meta-SS reported herein, and BTA 6 exhibited similarities with *iHS* reported by Qanbari et al. (2011) and Schwarzenbacher et al. (2012) in Brown Swiss and by Hayes et al. (2008) in Norwegian Red. Hayes et al. (2009) assessed evidence of divergent selection in Holstein and Angus using F_{ST} and *iHS*, and were able to detect signals in Holstein BTA 6 that resembled the *meta-SS* pattern found in the BSW dataset used in the present work. Moreover, Flori et al. (2009) examined F_{ST} within and across three French dairy cattle breeds, finding putative regions under selection that overlap the findings on BSW chromosomes 5 and 6 in the present study.

Based on the findings presented, the combination of multiple methods and the functional annotation strategies adopted seemed to be highly informative. Notwithstanding, some challenges still need to be overcome when considering scanning genome-wide data for selection sweeps. First, as similar genomic patterns can be produced by other phenomena, such as genetic drift, separating false

positives from real selection signals may not be trivial. Second, identified candidate regions often lacked spatial resolution, spanning from hundreds of kilobases to few megabases and comprising many genes. Third, distinguishing causal variants from nearby neutral loci may be the most difficult issue, as those variants were probably seldom typed in SNP arrays, and even with whole genome sequence data, variants in LD with the actual selected locus could have produced similar signals due to genetic hitch-hiking. Integrating different methodologies may help mitigating these problems, and should provide a valuable tool for seeking loci that are likely to have undergone recent artificial selection.

Finally, hypothesis making research implies proposing the function given the loci. In the present paper, this has meant inferring the source of selection for a given set of significant signals by extracting known gene functions and interaction information from available databases resources. Although the adopted functional annotation workflow using automated database mining and networking seemed to be a useful tool for providing insights on the driving forces behind the signals, the comprehensive nature of the annotation approach was expected to retrieve analysis artifacts due to systematic biases. Thus, hypothesis-driven investigations on the findings herein reported will contribute to elucidate which functions did undergo selection.

6. Acknowledgments

The authors wish to thank the U.S. Department of Agriculture, the University of Natural Resources and Life Sciences Vienna, and The HapMap and The Zebu Genome Consortia for providing the genotypes used in this paper. We want to express our highest gratitude to the European Science Foundation and the Advances in Farm Animal Genomic Resources project for supporting this research. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the authors or their respective institutions.

7. Competing interest

The authors have declared that no competing interests exist.

8. Financial disclosure

This research received support from the European Science Foundation and the Advances in Farm Animal Genomic Resources project (process nº 3726), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (processes 560922/2010-8 and 483590/2010-0), Fundacão de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processes 2011/16643-2 and 2010/52030-2) and USDA Agricultural Research Service (project 1265-31000-098D). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

9. Author's contributions

Conceived and designed the experiments: J. Sölkner, J. F. Garcia, Y. T. Utsunomiya, A. M. Pérez O'Brien. Performed the experiments: Y. T. Utsunomiya, A. M. Pérez O'Brien, G. Mészarós, A. S. Carmo. Analyzed the data: Y. T. Utsunomiya, A. M. Pérez O'Brien, G. Mészarós, A. S. Carmo. Contributed reagents/materials/analysis tools: J. F. Garcia, T. S. Sonstegard, C. P. Van Tassell, J. Sölkner. Wrote the paper: Y. T. Utsunomiya, J. Sölkner, J. F. Garcia, T. S. Sonstegard, G. Mészarós, C. P. Van Tassell, A. M. Pérez O'Brien, A. S. Carmo.

10. References

Ajmone-Marsan, P.; Garcia, J. F.; Lenstra, J. A. The Globaldiv Consortium On the Origin of Cattle: How Aurochs Became Cattle and Colonized the World. **Evolutionary Anthropology**, v. 19, p. 148-157, 2010.

Andersen, K. G.; Shylakhter, I.; Tabrizi, S.; Grossman, S. R.; Happi, C. T.; Sabeti, P. C. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 367, p. 868-877, 2012.

Arnandis, T.; Ferrer-Vicens, I.; García-Trevijano, E. R.; Miralles, V. J.; García, C.; Torres, L.; Viña, J. R.; Zaragozá, R. Calpains mediate epithelial-cell death during mammary gland involution: mitochondria and lysosomal destabilization. **Cell Death and Differentiation**, v. 19, n. 9, p. 1536-48, 2012.

Bionaz, M.; Loor, J. J. Gene Networks Driving Bovine Mammary Protein Synthesis During the Lactation Cycle. **Bioinformatics and Biology Insights**, v. 5, p. 83–98, 2011.

Bode, L.; Beermann, C.; Mank, M.; Kohn, G.; Boehm, G. Human and Bovine Milk Gangliosides Differ in Their Fatty Acid Composition. **The Journal of Nutrition**, v. 134, n. 11, p. 3016-3020, 2004.

Brann, D. W.; Mahesh, V. B. Excitatory amino acids: evidence for a role in the control of reproduction and anterior pituitary hormone secretion. **Endocrine Reviews**, v. 18, p. 678-700, 1997.

Decatur, W. A.; Fournier, M. J. rRNA modifications and ribosome function. **Trends Biochemical Sciences**, v. 27, p. 344-351, 2012.

Fay, J. C.; Wu, I. Hitchhiking under positive Darwinian selection. **Genetics**, v. 155, p. 1405-1413, 2000.

Flori, L.; Fritz, S.; Jaffrézic, F.; Boussaha, M.; Gut, I.; Heath, S.; Foulley, J.L.; Gautier, M. The genome response to artificial selection: a case study in dairy cattle. **PLoS ONE**, 4:e6595, 2009.

Fu, Y. X.; Li, W. H. Statistical tests of neutrality of mutations. **Genetics**, v. 133, p. 693-709, 1993.

Gadient, R. A.; Patterson, P. H. Leukemia inhibitory factor, interleukin 6, and other cytokines using the GP130 transducing receptor: roles in inflammation and injury. **Stem Cells**, v. 17, n. 3, p. 127-137, 1999.

Gautier, M.; Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. **Bioinformatics**, v. 28, n. 8, p. 1176-1177, 2012.

Grossman, S. R.; Shlyakhter, I.; Karlsson, E. K.; Byrne, E. H.; Morales, S.; Frieden, G.; Hostetter, E.; Angelino, E.; Garber, M.; Zuk, O.; Lander, E. S.; Schaffner, S. F.; Sabeti, P. C. A composite of multiple signals distinguishes causal variants in regions of positive selection. **Science**, v. 327, p. 883-886, 2010.

Grosz, M. D.; Macneil, M. D. Brief communication. The 'spotted' locus maps to bovine chromosome 6 in Hereford-cross population. **The Journal of Heredity**, v. 90, p. 233-236, 1999.

Hayes, B. J.; Chamberlain, A. J.; Maceachern, S.; Savin, K.; McPartlan, H.; MacLeod, I.; Sethuraman, L.; Goddard, M. E. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. **Animal Genetics**, v. 40, n. 2, p. 176-184, 2009.

Hayes, B. J.; Lien, S.; Nilsen, H.; Olsen, H. G.; Berg, P.; Maceachern, S.; Potter, S.; Meuwissen, T. H. The origin of selection signatures on bovine chromosome 6. **Animal Genetics**, v. 39, p. 105-111, 2008.

Huang D. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. **Nucleic Acids Research**, v. 37, n. 1, p. 1-13, 2009a.

Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. **Nature Protocols**, v. 4, n. 1, p. 44-57, 2009b.

Kinsella, R. J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; Kersey, P.; Flicek, P. Ensembl

BioMarts: a hub for data retrieval across taxonomic space. **Database (Oxford),** Bar030, 2011.

Kiss, T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. **Cell**, v. 109, p. 145-148, 2002.

Klungland, H. D.; Vage, I.; Gomez-Raya, L.; Adalsteinsson, S.; Lien, S. The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. **Mammalian Genome**, v. 6, p. 636–639, 1995.

Koohmaraie, M.; Geesink, G. H. Contribution of postmortem muscle biochemistry to the delivery of consistent meat quality with particular focus on the calpain system. **Meat Science**, v. 74, p. 34-43, 2006.

Kuhla, B.; Nürnberg, G.; Albrecht, D.; Görs, S.; Hammon, H. M.; Metges, C. C. Involvement of skeletal muscle protein, glycogen, and fat metabolism in the adaptation on early lactation of dairy cows. **Journal of Proteome Research**, v. 10, n. 9, p. 4252-62, 2011.

Matukumalli, L. K.; Lawley, C. T.; Schnabel, R. D.; Taylor, J. F.; Allan, M. F.; Heaton, M. P.; O'Connell, J.; Moore, S. S.; Smith, T. P.; Sonstegard, T. S.; Van Tassell, C. P. Development and characterization of a high density SNP genotyping assay for cattle. **PLoS One**, 4:e5350, 2009.

Merico, D.; Isserlin, R.; Stueker, O.; Emili, A.; Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. **PLoS One**, 5:e13984, 2010.

Momoeda, M.; Fukuta, S.; Iwamori, Y.; Taketani, Y.; Iwamori, M. Prolactin-dependent Expression of GD1α Ganglioside, as a Component of Milk Fat Globule, in the Murine Mammary Glands. **The Journal of Biochemistry**, v. 142, n. 4, p. 525-531, 2007.

Murray C.; Huerta-Sanchez, E.; Casey, F.; Bradley, D. G. Cattle demographic history modelled from autosomal sequence variation. **Philosophical Transactions of the Royal Society B**. v. 365, p. 2531-2539, 2010.

Nielsen, R.; Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. **Genetics**, v. 148, p. 929-936, 1998.

Oleksyk, T. K.; Smith, M. W.; O'Brien, S. J. Genome-wide scans for footprints of natural selection. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 365, p. 185-205, 2010.

Oleksyk, T. K.; Zhao, K.; De, L. a.; Gilbert, D. A.; O'Brien, S. J.; Smith, M. W. Identifying Selected Regions from Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two Human Populations. **PLoS ONE**, 3:e1712, 2008.

Ong, R. T-H., Teo, Y. Y. varLD: a program for quantifying variation in linkage disequilibrium patterns between populations. **Bioinformatics**, v. 26, n. 9, p. 1269-1270, 2010.

Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M. A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P. I.; Daly, M. J.; Sham, P. C. PLINK: a toolset for whole-genome association and population-based linkage analysis. **American Journal of Human Genetics**, v. 81, n. 3, p. 559-575, 2007.

Qanbari, S.; Gianola, D.; Hayes, B.; Schenkel, F.; Miller, S.; Moore, S.; Thaller, G.; Simianer, H. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. **BMC Genomics**, 12:318, 2011.

Quinlan, A. R.; Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, p. 841-842, 2010.

Rubin, C. J.; Zody, M. C.; Eriksson, J.; Meadows, J. R.; Sherwood, E.; Webster, M. T.; Jiang, L.; Ingman, M.; Sharpe, T.; Ka, S.; Hallböök, F.; Besnier, F.; Carlborg, O.; Bed'hom, B.; Tixier-Boichard, M.; Jensen, P.; Siegel, P.; Lindblad-Toh, K.; Andersson, L. Whole-genome resequencing reveals loci under selection during chicken domestication. **Nature**, v. 464, p. 587-591, 2010.

Sabeti, P. C.; Reich, D. E.; Higgins, J. M.; Levine, H. Z.; Richter, D. J.; Schaffner, S. F.; Gabriel, S. B.; Platko, J. V.; Patterson, N. J.; McDonald, G. J.; Ackerman, H. C.; Campbell, S. J.; Altshuler, D.; Cooper, R.; Kwiatkowski, D.; Ward, R.; Lander, E. S.

Detecting recent positive selection in the human genome from haplotype structure. **Nature**, v. 419, p. 832-837, 2002.

Sabeti, P. C.; Varilly, P.; Fry, B.; Lohmueller, J.; Hostetter, E.; Cotsapas, C.; Xie, X.; Byrne, E. H.; McCarroll, S. A.; Gaudet, R.; Schaffner, S. F.; Lander, E. S. ; International HapMap Consortium; Frazer, K. A.; Ballinger, D. G.; Cox, D. R.; Hinds, D. A.; Stuve, L. L.; Gibbs, R. A.; Belmont, J. W.; Boudreau, A.; Hardenbol, P.; Leal, S. M.; Pasternak, S.; Wheeler, D. A.; Willis, T. D.; Yu, F.; Yang, H.; Zeng, C.; Gao, Y.; Hu, H.; Hu, W.; Li, C.; Lin, W.; Liu, S.; Pan, H.; Tang, X.; Wang, J.; Wang, W.; Yu, J.; Zhang, B.; Zhang, Q.; Zhao, H.; Zhao, H.; Zhou, J.; Gabriel, S. B.; Barry, R.; Blumenstiel, B.; Camargo, A.; Defelice, M.; Faggart, M.; Goyette, M.; Gupta, S.; Moore, J.; Nguyen, H.; Onofrio, R. C.; Parkin, M.; Roy, J.; Stahl, E.; Winchester, E.; Ziaugra, L.; Altshuler, D.; Shen, Y.; Yao, Z.; Huang, W.; Chu, X.; He, Y.; Jin, L.; Liu, Y.; Shen, Y.; Sun, W.; Wang, H.; Wang, Y.; Wang, Y.; Xiong, X.; Xu, L.; Waye, M. M.; Tsui, S. K.; Xue, H.; Wong, J. T.; Galver, L. M.; Fan, J. B.; Gunderson, K.; Murray, S. S.; Oliphant, A. R.; Chee, M. S.; Montpetit, A.; Chagnon, F.; Ferretti, V.; Leboeuf, M.; Olivier, J. F.; Phillips, M. S.; Roumy, S.; Sallée, C.; Verner, A.; Hudson, T. J.; Kwok, P. Y.; Cai, D.; Koboldt, D. C.; Miller, R. D.; Pawlikowska, L.; Taillon-Miller, P.; Xiao, M.; Tsui, L. C.; Mak, W.; Song, Y. Q.; Tam, P. K.; Nakamura, Y.; Kawaguchi, T.; Kitamoto, T.; Morizono, T.; Nagashima, A.; Ohnishi, Y.; Sekine, A.; Tanaka, T.; Tsunoda, T.; Deloukas, P.; Bird, C. P.; Delgado, M.; Dermitzakis, E. T.; Gwilliam, R.; Hunt, S.; Morrison, J.; Powell, D.; Stranger, B. E.; Whittaker, P.; Bentley, D. R.; Daly, M. J.; de Bakker, P. I,; Barrett, J.; Chretien, Y. R.; Maller, J.; McCarroll, S.; Patterson, N.; Pe'er, I.; Price, A.; Purcell, S.; Richter, D. J.; Sabeti, P.; Saxena, R.; Schaffner, S. F.; Sham, P. C.; Varilly, P.; Altshuler, D.; Stein, L. D.; Krishnan, L.; Smith, A. V.; Tello-Ruiz, M. K.; Thorisson, G. A.; Chakravarti, A.; Chen, P. E.; Cutler, D. J.; Kashuk, C. S.; Lin, S.; Abecasis, G. R.; Guan, W.; Li, Y.; Munro, H. M.; Qin, Z. S.; Thomas, D. J.; McVean, G.; Auton, A.; Bottolo, L.; Cardin, N.; Eyheramendy, S.; Freeman, C.; Marchini, J.; Myers, S.; Spencer, C.; Stephens, M.; Donnelly, P.; Cardon, L. R.; Clarke, G.; Evans, D. M.; Morris, A. P.; Weir, B. S.; Tsunoda, T.; Johnson, T. A.; Mullikin, J. C.; Sherry, S. T.; Feolo, M.; Skol, A.; Zhang, H.; Zeng, C.; Zhao, H.; Matsuda, I.; Fukushima, Y.; Macer, D. R.; Suda, E.; Rotimi, C. N.; Adebamowo, C. A.; Ajayi, I.; Aniagwu, T.; Marshall, P. A.; Nkwodimmah, C.; Royal, C. D.; Leppert, M. F.; Dixon, M.; Peiffer, A.; Qiu, R.; Kent, A.; Kato, K.; Niikawa, N.; Adewole, I. F.; Knoppers, B. M.; Foster, M. W.; Clayton, E. W.; Watkin, J.; Gibbs, R. A.; Belmont, J. W.; Muzny, D.; Nazareth, L.; Sodergren, E.; Weinstock, G. M.; Wheeler, D. A.; Yakub, I.; Gabriel, S. B.; Onofrio, R. C.; Richter, D. J.; Ziaugra, L.; Birren, B. W.; Daly, M. J.; Altshuler, D.; Wilson, R. K.; Fulton, L. L.; Rogers, J.; Burton, J.; Carter, N. P.; Clee, C. M.; Griffiths, M.; Jones, M. C.; McLay, K.; Plumb, R. W.; Ross, M. T.; Sims, S. K.; Willey, D. L.; Chen, Z.; Han, H.; Kang, L.; Godbout, M.; Wallenburg, J. C.; L'Archevêque, P.; Bellemare, G.; Saeki, K.; Wang, H.; An, D.; Fu,

H.; Li, Q.; Wang, Z.; Wang, R.; Holden, A. L.; Brooks, L. D.; McEwen, J. E.; Guyer, M. S.; Wang, V. O.; Peterson, J. L.; Shi, M.; Spiegel, J.; Sung, L. M.; Zacharia, L. F.; Collins, F. S.; Kennedy, K.; Jamieson, R.; Stewart, J. Genome-wide detection and characterization of positive selection in human populations. **Nature**, v. 449, p. 913-918, 2007.

Scheet, P.; Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. **The American Journal of Human Genetics**, v. 78, p. 629-644, 2006.

Schwarzenbacher, H.; Dolezal, M.; Flisikowski, K.; Seefried, F.; Wurmser, C.; Schlötterer, C.; Fries, R. Combining evidence of selection with association analysis increases power to detect regions influencing complex traits in dairy cattle. **BMC Genomics**, 13:48, 2012.

Shi, Y.; Suh, Y. H.; Milstein, A. D.; Isozaki, K.; Schmid, S. M. Functional comparison of the effects of TARPs and cornichons on AMPA receptor trafficking and gating. **Proceedings of the National Academy of Sciences U.S.A.**, v. 107, p. 16315-16319, 2010.

Stella, A.; Ajmone-Marsan, P.; Lazzari, B.; Boettcher, P. Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. **Genetics**, v. 185, n. 4, p. 1451-1461, 2010.

Sugimoto, M.; Sasaki, S.; Watanabe, T.; Nishimura, S.; Ideta, A.; Yamazaki, M.; Matsuda, K.; Yuzaki, M.; Sakimura, K.; Aoyagi, Y.; Sugimoto, Y. Ionotropic glutamate receptor AMPA 1 is associated with ovulation rate. **PLoS One**, 5:e13817, 2010.

Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. **Nucleic Acids Research**, 39:D561-8, 2011.

Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics**, v. 123, p. 585-595, 1989.

Tang, K.; Thornton, K. R.; Stoneking, M. A new approach for using genome scans to detect recent positive selection in the human genome. **PLoS Biology**, 5:e171, 2007.

The Bovine Genome Sequencing and Analysis Consortium; Elsik, C. G.; Tellam, R. L.; Worley, K. C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, v. 324, n. 5926, p. 522-528, 2009.

The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. **Science**, v. 324, n. 5926, p. 528-532, 2009.

Villa-Angulo, R.; Matukumalli, L. K.; Gill, C. A.; Choi, J.; Van Tassell, C. P..; Grefenstette, J. J. High-resolution haplotype block structure in the cattle genome. **BMC Genetics**, 10:19, 2009.

Voight, B. F.; Kudaravalli, S.; Wen, X.; Pritchard, J. K. A map of recent positive selection in the human genome. **PLoS Biology**, 4:e72, 2009.

Weir, B. S.; Cockerham, C. C. Estimating F-Statistics for the analysis of population structure. **Evolution**, v. 38, n. 6, p. 1358-1370, 1984.

Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. **Journal of Evolutionary Biology**, v. 18, p. 1368-1373, 2005.

Wilde, C. J.; Addey, C. V. P.; Li, P.; Fernig, D. G. Programmed cell death in bovine mammary tissue during lactation and involution. **Experimental Physiology**, v. 82, p. 943-953, 1997.

Williamson, S. H.; Hubisz, M. J.; Clark, A. G.; Payseur, B. A.; Bustamante, C. D.; Nielsen, R. Localizing recent adaptive evolution in the human genome. **PLoS Genetics**, 3 e90, 2007.

Zhao, C.; Tian, F.; Yu, Y.; Luo, J.; Mitra, A.; Zhan, F.; Hou, Y.; Liu, G.; Zan, L.; Updike, M. S.; Song, J. Functional Genomic Analysis of Variation on Beef

Tenderness Induced by Acute Stress in Angus Cattle. **Comparative and Functional Genomics**, 756284, 2012.

Zimin, A. V.; Delcher, A. L.; Florea, L.; Kelley, D. R.; Schatz, M. C.; Puiu, D.; Hanrahan, F.; Pertea, G.; Van Tassell, C. P..; Sonstegard, T. S.; Marçais, G.; Roberts, M.; Subramanian, P.; Yorke, J. A.; Salzberg, S. L. A whole-genome assembly of the domestic cow, *Bos taurus*. **Genome Biology**, 10:R42, 2009.

Zoico, E.; Roubenoff, R. The role of cytokines in regulating protein metabolism and muscle function. **Nutrition Reviews**, v. 60, n. 2, p. 39-51, 2002.

APPENDICES

APPENDIX A - Extended methods for weighted FASTA

1. The standard FASTA

Variance-components models are the gold standard for genome-wide association analysis of single nucleotide polymorphism (SNP) markers accounting for relatedness and population substructure. However, fitting all model parameters for every tested SNP makes the process computationally demanding. In order to overcome this problem, approximation approaches have been proposed, which divide the estimation of parameters into two steps: first, a variance-components model is fitted to the data; then, the significance of each marker is either obtained from score tests corrected for the variance-covariance matrix (CHEN & ABECASIS, 2007; ZHANG et al., 2010; KANG et al., 2010; LIPPERT et al., 2011) or least squares regressions using residuals as the dependent variable (AULCHENKO et al., 2007; AMIN et al., 2007).

The *Fast Association Score Test-based Analysis* (FASTA) method (CHEN & ABECASIS, 2007) comprises fitting a variance-components model to the data in order to obtain the variance-covariance matrix for the phenotypes, which is then used to compute allele substitution effects for each tested SNP. The variance-components model is based on the polygenic model:

$$y = X\beta + u + \varepsilon \quad \textbf{[1]}$$

where y is the vector of phenotypes observed for n individuals, X is a nx k design matrix of k covariates, β is a column vector of size k of fixed effects of covariates, and u and ε are vectors of unobserved random additive genetic and residual effects, respectively.

The polygenic model involves partitioning the total trait variance σ_T^2 in two components: variance due to genetic differences among individuals, namely additive genetic variance σ_u^2, and a residual variance σ_ε^2. Additive genetic effects are assumed to follow a $MVN(0, \Phi\sigma_u^2)$, where Φ is a relationship matrix. The diagonal elements of matrix Φ are individual variances, which can be expressed as $1 + F$,

where F is the inbreeding coefficient (ASTLE & BALDING, 2009). The off-diagonal elements of this matrix are covariances between individuals, which can be expressed as twice their kinship coefficient. Unbiased estimates of kinship coefficients between individuals can be obtained from genotypic data as (AMIN et al., 2007; ASTLE & BALDING, 2009):

$$\hat{f}_{i,j} = \frac{1}{L}\sum_{l=1}^{L} \frac{(g_{l,i}-p_l)(g_{l,j}-p_l)}{p_l(1-p_l)} \textbf{ [2]}$$

where $\hat{f}_{i,j}$ is the estimated genomic kinship between individuals i and j, L is the total number of loci used for the calculation, p_l is the reference allele frequency for locus l, and $g_{l,i}$ and $g_{l,j}$ are the locus l genotypes for individuals i and j, respectively (coded as 0, 1 or 2 reference alleles).

The model specifies that the random residual effect for each individual is normally distributed with mean zero and variance σ_ε^2. These residuals are assumed to be independent between individuals, and the joint distribution of residuals is defined as $MVN(0, I\sigma_\varepsilon^2)$, where I is an identity matrix. In this setting, variances are assumed to be equal among individual phenotypes, and the variance-covariance matrix is defined as:

$$\Omega = \Phi\sigma_u^2 + I\sigma_\varepsilon^2 \textbf{ [3]}$$

The log-likelihood function of the model is then specified as:

$$log(L) = -\frac{n}{2}log(2\pi) - \frac{1}{2}\,log|\Omega| - \frac{1}{2}(y - X\beta)^T\Omega^{-1}(y - X\beta) \textbf{ [4]}$$

Maximum likelihood estimates for each model parameter are obtained from this function by using an optimization algorithm. Next, the estimated σ_u^2 and σ_ε^2 are used in **[3]** to calculate the variance-covariance matrix at the point of maximum likelihood. Then, for each SNP, the allele substitution effect and its variance are obtained using generalized least squares equations:

$$\hat{\beta}_{SNP} = \left(\tilde{g}'\Omega^{-1}\tilde{g}\right)^{-1}\tilde{g}'\Omega^{-1}\tilde{y} \textbf{ [5]}$$

$$var\left(\hat{\beta}_{SNP}\right) = \left(\tilde{g}'\Omega^{-1}\tilde{g}\right)^{-1} \textbf{ [6]}$$

Where $\hat{\beta}_{SNP}$ and $var(\hat{\beta}_{SNP})$ are the estimated allele substitution effect and its variance, respectively; $\tilde{g} = g - E(g)$, where g is the vector of observed genotypes for a given marker coded as 0, 1 or 2 reference alleles, and $E(g)$ is a genotype mean;

and $\tilde{y}$ is the vector of dependent phenotype residuals $\tilde{y} = y - X\hat{\beta}$ (phenotype adjusted for the estimated fixed effects).

2. Adapting FASTA to account for heterogeneity of variance in dEBVs

Following Garrick et al. (2009), a deregressed estimated breeding value (dEBV) represents a pseudo-phenotype that summarizes all the information available on the individual and its relatives, as if it was a single observation. In this special case, variances are unequal among individual pseudo-phenotypes as the number of repeated measures or observations of relatives for each individual varies. Thus, it is necessary to adapt the polygenic model to allow for heterogeneity of variance among individuals. This can be achieved by replacing the diagonal of the identity matrix I in the residual variance matrix by individual weights that are proportional to the estimate errors. Now, residual random effects are assumed to follow a $MVN(0, R\sigma_\varepsilon^2)$, where R is a diagonal weight matrix.

The main objective in maximizing the likelihood in **[4]** is to obtain estimates for the variance components and random/fixed effects of the model. However, it is easy to show that the polygenic model in FASTA is equivalent to the animal model:

$$y = X\beta + Zu + \varepsilon \quad \textbf{[7]}$$

where Z is a design matrix of order nx n that relates phenotypes to random animal effects. The random effects are assumed to be distributed as:

$$\begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Phi\sigma_u^2 & 0 \\ 0 & R\sigma_\varepsilon^2 \end{bmatrix}\right) \quad \textbf{[8]}$$

This model can be fitted using the mixed model equations (MME) developed by Henderson (1973):

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + \Phi^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \quad \textbf{[9]}$$

Where λ is the variance ratio $\sigma_\varepsilon^2/\sigma_u^2$. Now, variance components can be estimated via restricted (residual) maximum likelihood (REML), and $\hat{\beta}$ and $\hat{u}$ can be simultaneously estimated. The variance-covariance matrix for the phenotypes is then defined as $V = Z'(\Phi\sigma_u^2)Z + R\sigma_\varepsilon^2$. As dEBVs are single observations, matrix Z is an identity

matrix, so the model in **[7]** can be rewritten as $y = X\beta + u + \varepsilon$, and the variance-covariance matrix becomes

$$V = \Phi\sigma_u^2 + R\sigma_\varepsilon^2 \quad \textbf{[10]}$$

Note that $R = I$ when variances are equal among individual phenotypes, in which case $V = \Omega = \Phi\sigma_u^2 + I\sigma_\varepsilon^2$, and the animal model is equivalent to the original polygenic model used in FASTA. Thus, we used REML to estimate variance components and MME to estimate fixed and random effects in the first step of FASTA, allowing for fitting phenotypes with unequal variances. The allele substitution effect and the variance of each SNP were then obtained as in the original FASTA approach, following **[5]** and **[6]**.

3. Choosing appropriate weights for dEBVs

The choice of weights to be used can vary according to the nature of the response variable being analyzed. In the case of dEBVs, Garrick et al. (2009) proposed the following weights to account for heterogeneity of variance:

$$w_i = \frac{(1-h^2)}{\left[\left(c+\frac{1-r_i^2}{r_i^2}\right)h^2\right]} \quad \textbf{[11]}$$

where c is the assumed proportion of the genetic variance not explained by markers due to partial genome coverage and incomplete linkage disequilibrium between markers and causal variants, h^2 is the estimated heritability of EBVs before deregression, and r_i^2 is the reliability of the dEBV of animal i. These weights are diagonal elements of the inverse of the scaled residual variance matrix $R\sigma_\varepsilon^2$, with σ_ε^2 being factored out before inversion. Hence, the weight matrix R can be expressed in the non-inverted form as:

$$R = \begin{bmatrix} 1/w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/w_i \end{bmatrix} \quad \textbf{[12]}$$

Garrick et al. (2009) argued that the choice of a value for c can be made by assessing a range of values or by estimating c from validation analyses. In practice, they showed that the impact of the assumed value of c is to influence the relative value of individuals with accurate information, in comparison to individuals with less

reliable information. When $c = 0$, weighting by w is equivalent to weighting observations by the inverse of their variances, which in the case of dEBVs can be approximated by $1/(1 - r^2)$. As the use of too large a value of c would result in little contrast between dEBVs with low and high accuracy, and the use of too small a value of c would result in excessive emphasis on dEBVs with high accuracy, we decided to fix c at 0.5.

4. References

Amin, N.; Van DuijnM C. M.; Aulchenko, Y. S. A genomic background based method for association analysis in related individuals. **PLoS One**, 2:e1274, 2007.

Astle, W.; Balding, D.J. Population structure and cryptic relatedness in genetic association studies. **Statistical Science**, v. 24, p. 451-71, 2009.

Aulchenko, Y. S.; de Koning, D-J.; Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. **Genetics**, v. 177, p. 577-585, 2007.

Chen, W. M.; Abecasis, G. R. Family-based association tests for genomewide association scans. **American Journal of Human Genetics**, v. 81, p. 913-26, 2007.

Garrick, D. J.; Taylor, J. F.; Fernando, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, 41:55, 2009.

Henderson CR. Sire evaluation and genetic trends. **Journal of Animal Science**, p. 10-41, 1973.

Kang, H. M.; Sul, J. H.; Service, S. K.; Zaitlen, N. A.; Kong, S. Y; Freimer, N. B.; Sabatti, C.; Eskin, E. Variance component model to account for sample structure in genome-wide association studies. **Nature Genetics**, v. 42, 348-54, 2010.

Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C. M. Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. **Nature Methods**, v. 8, p. 833-835, 2011.

Zhang, Z.; Ersoz, E.; Lai, C. Q.; Todhunter, R. J.; Tiwari, H. K.; Gore, M. A.; Bradbury, P. J.; Yu, J.; Arnett, D. K.; Ordovas, J. M.; Buckler, E. S. Mixed linear model approach adapted for genome-wide association studies. **Nature Genetics,** v. 42, p. 355-360, 2010.

APPENDIX B - Cryptic relatedness control and functional annotation

1. Cryptic relatedness control

We were interested in account for as much diversity as possible and keep only unrelated individuals within our dataset. It was known that BOKU and ZGC samples lodge different degrees of cryptic relatedness. According to The Bovine HapMap Consortium (2009), their sampling strategy involved genotyping individuals that were unrelated for ≥ 4 generations, but each breed had at least one sire, dam and progeny trio. We expected to find duplicates within HapMap samples, because we knew the consortium had genotyped some animals twice for genotype quality assessment. It was also possible that there were duplicated samples between BOKU and HapMap or ZGC and HapMap, because high ranked sires may have been genotyped by the three initiatives. As our access to pedigree information was limited, we investigated our dataset for pairwise allele identity using *PLINK*. The method adopted is based on Identity by Descent (IBD) and was described by Purcell and collaborators (2007). Briefly, it uses a method-of-moments approach to estimate the probability of sharing 0, 1 or 2 alleles identical by descent for any two individuals, assuming they come from the same homogeneous, random-mating population. If we denote IBS states as *I* and IBD states as *Z* (in both cases, the possible states being 0, 1, and 2), then we can express the prior probability of IBS sharing as:

$$P(I = i) = \sum_{z=0}^{z=i} P(I = i \mid Z = z)P(Z = z)$$

As described in detail in Purcell et al (2007), for each SNP, the $P(I \mid Z)$ is specified in terms of the allele frequency; averaging over all SNPs, we obtain the expected global

value for $P(I \mid Z)$. Then, rearranging the three equations implied by the equation above, we solve for $P(Z = 0)$, $P(Z = 1)$, and $P(Z = 2)$ and calculate:

$$\hat{\pi} = \frac{P(Z = 1)}{2} + P(Z = 2)$$

which is an estimation for the proportion of alleles shared identically by descent. Expected values of $\hat{\pi}$ for the different types of relationship can be found in Table 1B. This estimate and its expected values were used for investigation of cryptic relatedness within breeds.

First, we looked for possible replicates. Duplicated (or even monozygote twins) samples usually present P(Z = 2) ~ 1 and $\hat{\pi}$ ~ 1 (i.e. all alleles are identical by descent). Thus, pairs of samples showing $\hat{\pi} \geq 0.9$ were considered duplicates, and one of the samples was randomly excluded. Second, heatmaps of $\hat{\pi}$ values were drawn for each breed, in order to obtain an overall view of the cryptic relatedness present within the datasets. To improve visualization even more, Euclidean distances were calculated from $\hat{\pi}$ dissimilarities and samples were clustered according to the amount of alleles shared by descent. Finally, potential Parent-Offspring and Full-Siblings pairs were considered confounders for our analyses and at least one sample was excluded for each pair identified. As a single sample can hold first degree relationship with one or more samples at the same time, it is clear that its exclusion would solve the confounding effects and preserve more samples within the dataset than excluding random members of each pair. Thus, we developed an algorithm (written in *R*) that performs conservative exclusion of samples. For a given cryptic relatedness threshold (in this case, $\hat{\pi} > 0.4$), the following procedure is executed:

a. For each sample, count the number of cryptic relationships it holds.
b. Sort samples by count score.
c. Exclude the sample with the highest score.
d. Repeat a, b and c until all scores are equal to 0.

Table 1B. Different types of relatedness and their IBD values

Type	Degree	E($\hat{\pi}$)	$P(Z = 0)$	$P(Z = 1)$	$P(Z = 2)$
Full-Sibling	1-2	0.5000	0.2500	0.5000	0.2500
Half-Sibling	2	0.2500	0.5000	0.5000	0.0000
Grandparent-grandchild	2	0.2500	0.5000	0.5000	0.0000
Avuncular	2-3	0.2500	0.5000	0.5000	0.0000
First-Cousin	3	0.1250	0.7500	0.2500	0.0000
Unrelated	-	0.0000	1.0000	0.0000	0.0000
Half-Avuncular	3	0.1250	0.7500	0.2500	0.0000
Half-First-Cousin	4	0.0625	0.8750	0.1250	0.0000
Half-Sib+First-Cousin	2-3	0.3750	0.3750	0.5000	0.1250
Parent-Offspring	1	0.5000	0.0000	1.0000	0.0000
MZ-Twins	0	1.0000	0.0000	0.0000	1.0000

We found 7 BOKU-HapMap, 3 ZGC-HapMap and 3 HapMap-HapMap duplicates. After removal of replicated samples, we plotted the $\hat{\pi}$ heatmap (Figure 1B) and carried out a principal coordinates analysis (Figure 2B) to check for the integrity of our genotype files and sample tracking. The number of remaining samples for each breed after duplicates and first degree relationship ($\hat{\pi}$ > 0.4) removal were: 24 ANG, 44 BSW (13 HapMap and 31 BOKU), 23 GYR and 581 NEL (24 HapMap and 557 ZGC). As NEL exhibited a sample size much larger than the other breeds, 45 individuals were sampled from the total 581 (all 24 remaining HapMap samples + 21 ramdom ZGC samples), in order to do fair comparisons. We decided to keep all possible HapMap genotypes because the sampling strategy adopted by the HapMap consortium attempted to account for as much within-breed diversity as possible.

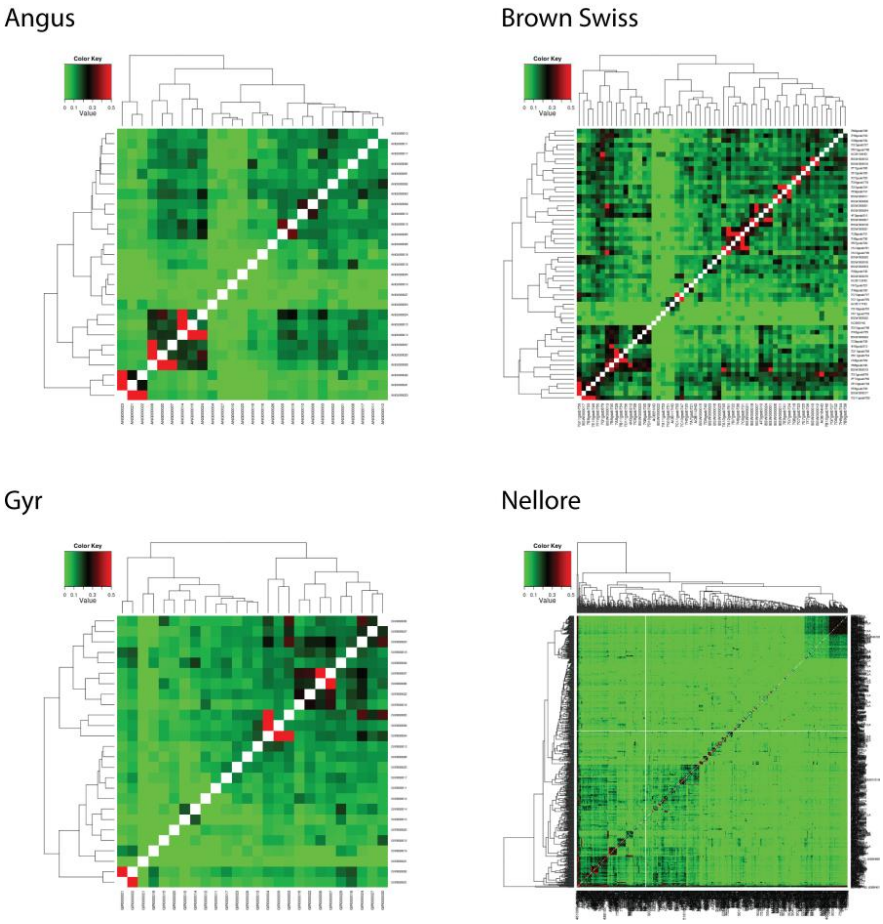

Figure 1B. Heatmap and clustering of samples based on relatedness, as measured by $\hat{\pi}$. Values range from 0 (green - completely unrelated samples) to 0.5 (red - IBD sharing of half of the alleles, corresponding to Parent-Offspring or Full-Siblings pairs).

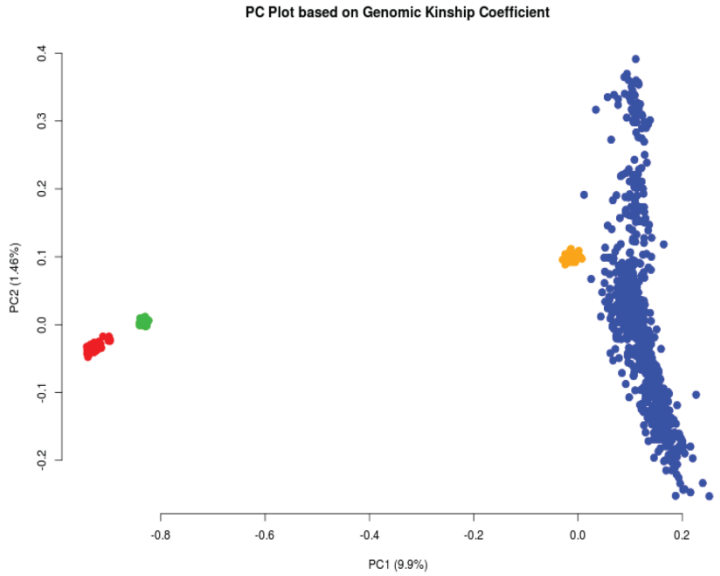

Figure 2B. Principal Coordinates Analysis. Red = BSW, Green = ANG, Orange = GYR and Blue = NEL. Percentages inside brackets correspond to proportion of variance explained by the respective eigenvectors.

2. Functional annotation

For any peak crossing the significance line, we applied three different strategies for the annotation of functional features.

Strategy 1: Since any given gene harboring signals is a direct candidate, the first approach consisted on checking if any significant SNP was intragenic via mining the *Ensembl Variation 67* database with the *Ensembl Biomart tool* (Kinsella *et al*., 2011).

Strategy 2: The closest gene in the vicinity of the most relevant SNP of a peak may be the responsible for the signal. Hence, the second strategy comprehended isolating the most significant SNP from each observed peak and mapping the closest gene to it. For that matter, we downloaded the Bovine UMD3.1 gene set from *Ensembl Genes 67* database via *Biomart tool* and used the *ClosestBed* algorithm from the *BedTools* software (Quinlan & Hall, 2010).

Strategy 3: There are cases where variants in multiple genes in linkage disequilibrium (LD) with the marker contribute to the signal together, because functionally related genes are often spatially close to each other. In fact, the usage of SNP chips is driven by the hypothesis that high marker density coverage of the genome is capable of capturing most of the genomic information by LD and haplotype structure. Thus, our third approach was a LD-based window scheme, divided into three steps.

Step 1: Every SNP crossing the significance line was defined as a 'core SNP'.

Step 2: We walked down to proximal and distal chromosome positions calculating correlations between the core SNP and the neighbor markers, checking if they tagged the core SNP or not based on r^2 values. The r^2 threshold adopted to declare that one marker tagged the core SNP was set to 0.7. The positions of the last tag markers on both sides of the core SNP, i.e., positions from where r^2 decayed below the defined threshold or the tag marker distance from the core position exceeded 1 Mb, were set as the boundaries of a window. This analysis was done in *PLINK*, using the options *--show-tags --list-all --tag-r2 0.7 --tag-kb 1000*.

Step 3: The retrieved window was interpreted as a single locus, and any gene overlapping it was considered to be in LD with the core SNP, thus a candidate for being involved with the selection signal. Such genes were therefore annotated. For the sake of marker density and region resolution, we used the lists of SNPs passing within breed QCs, regardless of ancestral allele information, instead of the unified list. For core SNPs where no window boundaries could be determined, we included the closest gene in the vicinity to the list. As some windows may also overlap, the derived gene list was then parsed to exclude repeated gene names and subsequently processed in *DAVID* (Huang *et al*., 2009a; Huang *et al*., 2009b) for annotation of functional terms. We used the default parameters for each breed gene list, pooling together all genes annotated across the genome to reveal over-represented functional terms. Our hypothesis was not that all genome signals detected came from a single selection event, but that some sweeps may have shared the same functional background, i.e. the same selection force. Therefore, genes from different regions and chromosomes may cluster together or not based on their function, revealing biological processes, rather than single genes, that undergone

selection. Finally, we used the *Enrichment Map Cytoscape plug-in* (Merico *et al*., 2010) to build networks of inter-related terms based on the number of overlapping genes. Terms were drawn as nodes (circles). Edges linking nodes represented gene sharing, and their thickness the degree of gene set overlap.

3. References

Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. **Nature Protocols**, v. 4, n. 1, p. 44-57, 2009a.

Huang, D. W.; Sherman, B.T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. **Nucleic Acids Research**, v. 37, n. 1, p. 1-13, 2009b.

Kinsella, R. J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; Kersey, P.; Flicek, P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. **Database (Oxford),** Bar030, 2011.

Merico, D.; Isserlin, R.; Stueker, O.; Emili, A.; Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. **PLoS One**, 5:e13984, 2010.

Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M. A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P. I.; Daly, M. J.; Sham, P. C. PLINK: a toolset for whole-genome association and population-based linkage analysis. **American Journal of Human Genetics**, v. 81, n. 3, p. 559-575, 2007.

Quinlan, A. R.; Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, p. 841-842, 2010.

The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. **Science**, v. 324, n. 5926, p. 528-532, 2009.

APPENDIX C - Supplementary figures

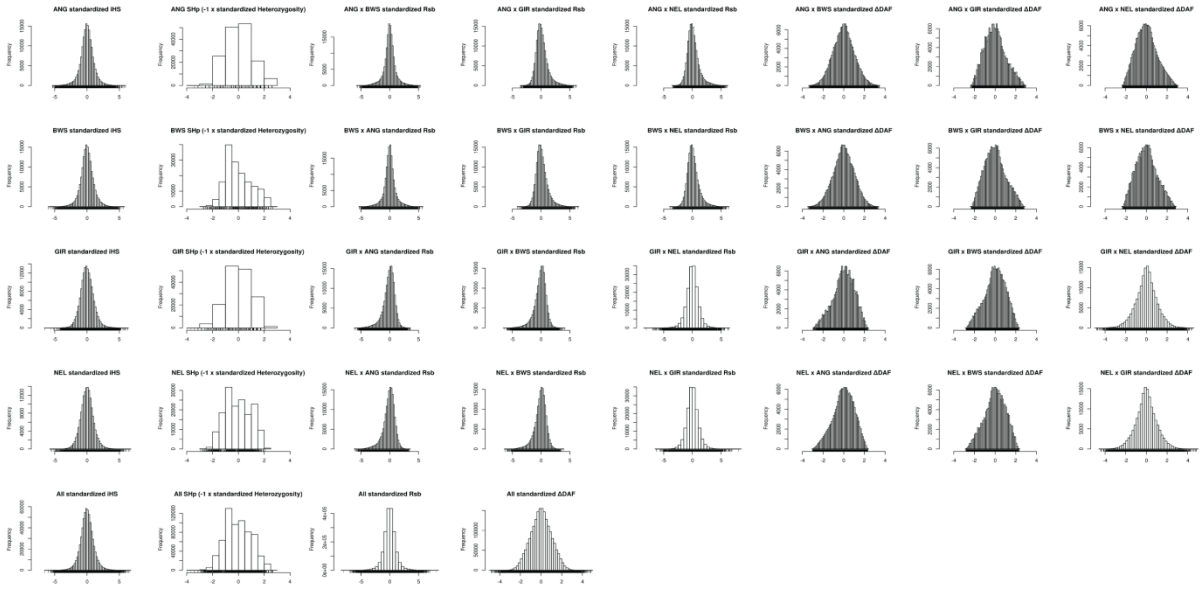

Figure 1C. Histogram for each individual standardized test score

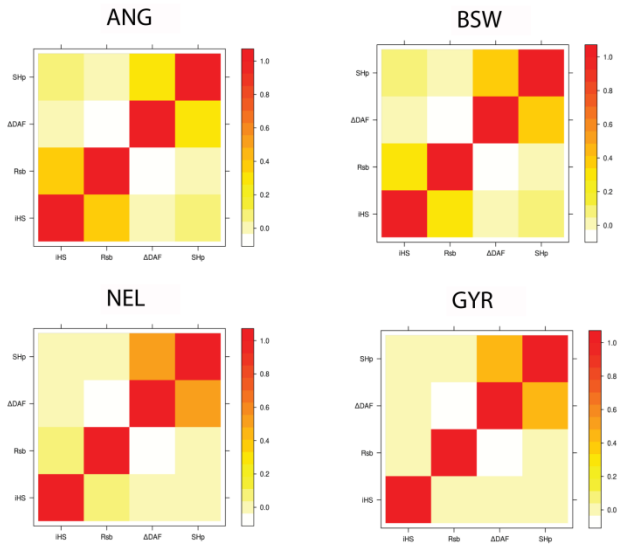

Figure 2C. Pearson correlations between each individual test Z-transformed *P*-values

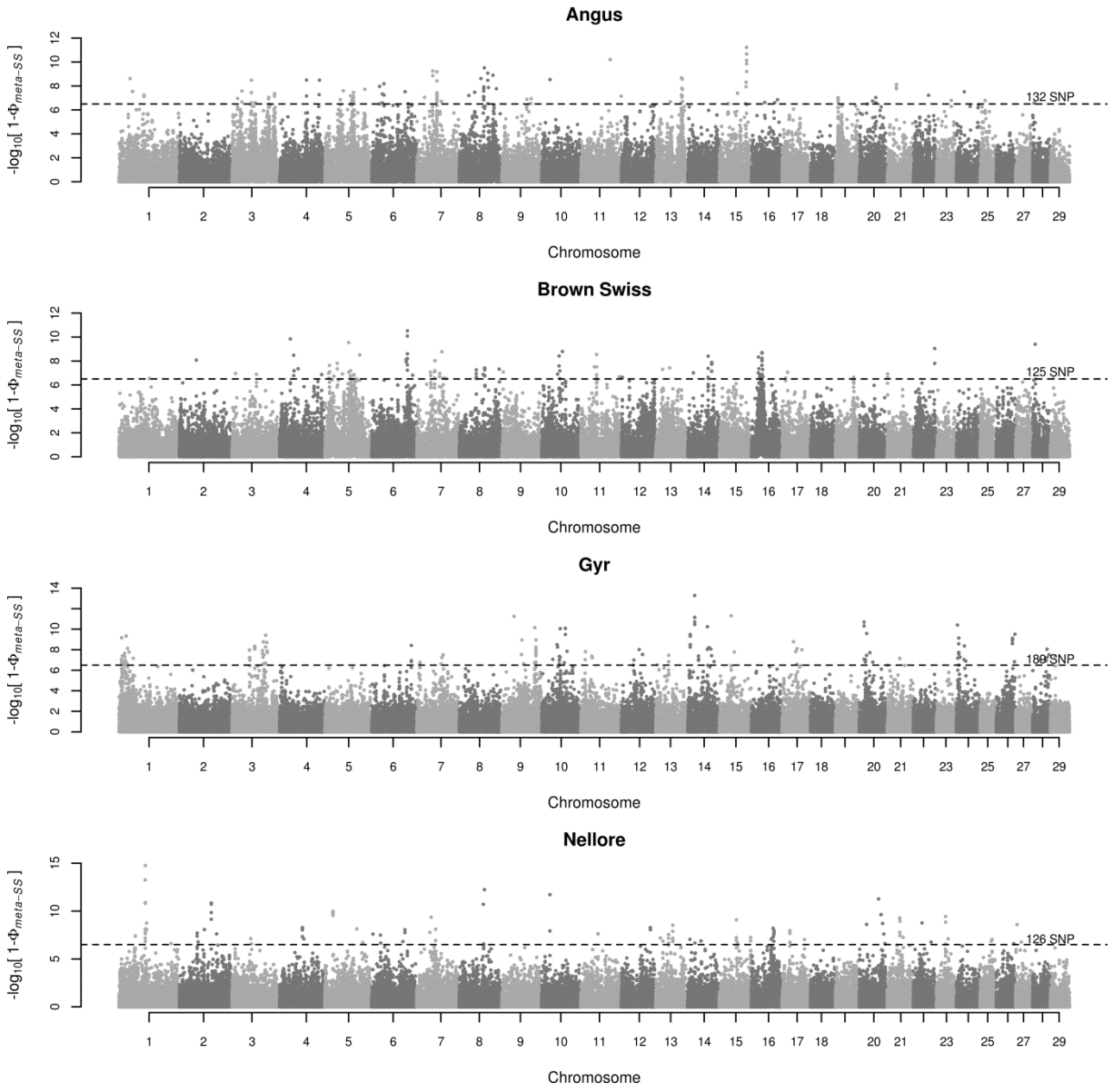

Figure 3C. Manhattan plots of genome-wide *meta-SS* -log10(*P*-values) combining within breeds tests only.

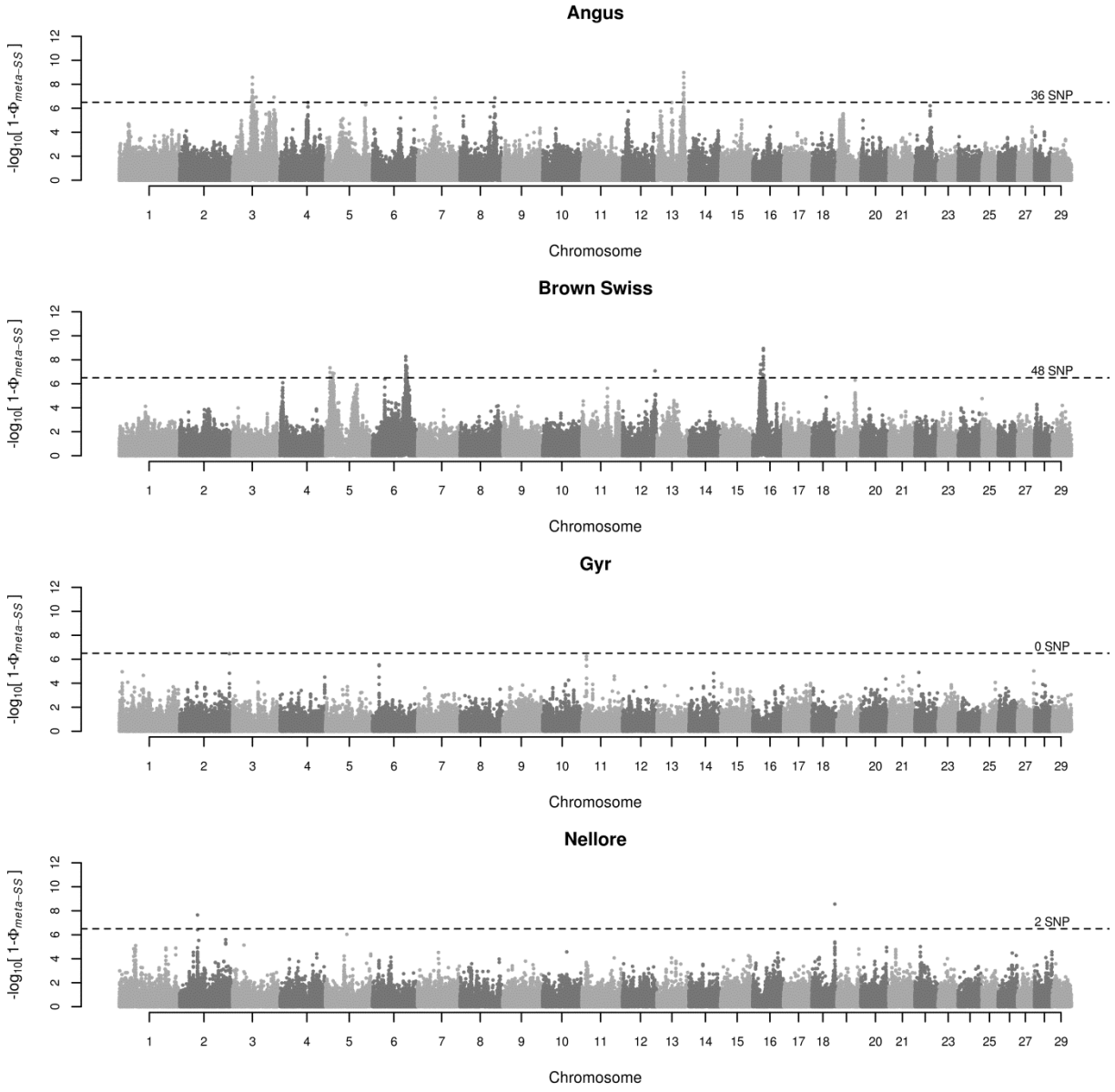

Figure 4C. Manhattan plots of genome-wide *meta-SS* -log10(*P*-values) combining between breeds tests only.