UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

ON THE ORIGIN AND SPREAD OF THE BOVINE *PLAG1* MUTATION

Yuri Tani Utsunomiya
Médico Veterináro

2017

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

ON THE ORIGIN AND SPREAD OF THE BOVINE *PLAG1* MUTATION

Yuri Tani Utsunomiya

Orientador: José Fernando Garcia

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

2017

Utsunomiya, Yuri Tani
U89o On the Origin and spread of the bovine PLAG1 mutation / Yuri Tani Utsunomiya. – – Jaboticabal, 2017
ix, 113 p. : il. ; 28 cm

Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2017
Orientador: José Fernando Garcia
Banca examinadora: Johann Sölkner, Flávia Lombardi Lopes, Paolo Ajmone-Marsan, Celso Luis Marino
Bibliografia

1. Cattle. 2. *PLAG1*. 3. Haplotype. 4. Pleiotropy. 5. Stature. 6. Reproduction. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 619:612.6:636.2

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

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ON THE ORIGIN AND SPREAD OF THE BOVINE *PLAG1* MUTATION

AUTOR: YURI TANI UTSUNOMIYA
ORIENTADOR: JOSE FERNANDO GARCIA

Aprovado como parte das exigências para obtenção do Título de Doutor em MEDICINA VETERINÁRIA, área: REPRODUÇÃO ANIMAL pela Comissão Examinadora:

Prof. Dr. JOSE FERNANDO GARCIA
Departamento de Apoio, Produção e Saude Animal / FMVA/UNESP - Araçatuba

Prof. Dr. JOHANN SÖLKNER
Departamento de Ciência Animal / Universität für Bodenkultur Wien (BOKU), Austria

Profa. Dra. FLÁVIA LOMBARDI LOPES
Departamento de Apoio, Produção e Saúde Animal / FMVA/UNESP - Araçatuba

Prof. Dr. PAOLO AJMONE MARSAN
Istituto di Zootecnica / Facoltà di SCIENZE AGRARIE, ALIMENTARI E AMBIENTALI / Italia

Prof. Dr. CELSO LUIS MARINO
Departamento de Genética / Instituto de Biociências de Botucatu - UNESP

Jaboticabal, 06 de dezembro de 2017

ABOUT THE AUTHOR

BSc in Veterinary Sciences (2010), São Paulo State University (Unesp), School of Veterinary Medicine, Araçatuba-SP Brazil. MSc in Veterinary Sciences (2013), São Paulo State University (Unesp), School of Agricultural and Veterinarian Sciences, Jaboticabal-SP Brazil. Co-author of 41 peer-reviewed articles, 4 book chapters and 2 open source software in collaboration with scientists from 10 countries.

“We are not to tell nature what she’s gotta be. She's always got better imagination than we have”

Richard Phillips Feynman

ACKNOWLEDGMENTS

To my family, for understanding and supporting my pull for Nature.

To Prof. José Fernando Garcia, who generously nurtured my passion for Science.

To Rafaela Beatriz Pintor Torrecilha, for being my best friend.

To all my dear friends from the Animal Biochemistry and Molecular Biology Laboratory, for all support and friendship in the past 10 years.

To Prof. Johann Sölkner and my friends from the University of Natural Resources and Life Sciences, for another wonderful year in the beautiful Vienna.

To the São Paulo Research Foundation (FAPESP) for financially supporting this research.

And finally, to the São Paulo State University (Unesp), the School of Veterinary Medicine from Araçatuba (FMVA) and the School of Agricultural and Veterinarian Sciences from Jaboticabal (FCAV), for providing me technical training and a fruitful scientific environment.

Yuri Tani Utsunomiya

Jaboticabal, December 2017

SUMMARY

SOBRE A ORIGEM E DISPERSÃO DA MUTAÇÃO DO GENE *PLAG1* EM BOVINOS

RESUMO - O gene 1 do adenoma pleomórfico (*PLAG1*) apresenta evidência de seleção positiva recente e associação com tamanho corporal e fertilidade em um grande número de raças bovinas ao redor do mundo. Tendo em vista sua recentemente descoberta função como fator de transcrição para o gene do fator de crescimento semelhante à insulina 2 (*IGF2*), o *PLAG1* possui papel emergente como um dos principais reguladores do crescimento e da reprodução em bovinos. Apesar de sua importância, a variante de sequência de DNA responsável pelos efeitos pleiotrópicos atribuídos ao *PLAG1* em bovinos permanece desconhecida. Também não está claro se a mesma mutação explica as associações fenótipo-genótipo encontradas em diferentes populações bovinas. Além disso, ainda é incerto onde e quando ocorreu a pressão de seleção responsável pelo aumento da frequência da mutação do *PLAG1*. No presente trabalho, reportamos o desenvolvimento de um pacote para o software estatístico R, o qual é direcionado à análise de haplótipos como preditores para variantes genéticas não observadas. Através da aplicação desta ferramenta a dados genômicos de bovinos oriundos de diversas regiões do mundo, encontramos evidência indicando que um único alelo derivado do *PLAG1* aumentou em frequência rapidamente em bovinos *Bos taurus* do noroeste europeu entre os séculos XVI e XVIII. Este período é reconhecido como a última onda de aumento de estatura em bovinos por meio de registros arqueológicos. Os dados também sugerem que o alelo foi introgredido em *B. taurus* não europeu e raças *Bos indicus* entre os séculos XIX e XX, adquirindo uma distribuição quase global no último século. Análises de DNA antigo revelaram que esta mutação segrega em gado do noroeste europeu há pelo menos 1.000 anos. Em conjunto, estes resultados implicam um papel central da mutação do *PLAG1* em recentes mudanças de tamanho corporal em bovinos.

Palavras-chave: Bovinos; *PLAG1*; Haplótipo; Pleiotropia; Estatura; Reprodução

ON THE ORIGIN AND SPREAD OF THE BOVINE *PLAG1* MUTATION

ABSTRACT - The pleomorphic adenoma gene 1 (*PLAG1*) presents both evidence of recent positive selection and association with body size and fertility in a wide range of worldwide cattle breeds. Considering its recently uncovered function as a transcription factor for the insulin-like growth factor 2 gene (*IGF2*), *PLAG1* is emerging as a major regulator of bovine growth and reproduction. In spite of its importance, the causal DNA sequence variant underlying the pleiotropic effects of *PLAG1* in cattle remains unknown. It is also unclear whether the same mutation accounts for the phenotype-genotype associations detected across different cattle populations. Furthermore, when and where the selective pressure responsible for increasing the frequency of the *PLAG1* mutation occurred is still uncertain. Here, we report the development of a package for the R statistical software to analyze haplotypes as surrogates for unobserved genetic variants. By applying this tool to genomic data of worldwide cattle breeds, we found evidence that a single bovine *PLAG1* derived allele increased rapidly in frequency in Northwestern European *Bos taurus* populations between the 16th and 18th centuries. This period is recognized as the last wave of increase in bovine stature from archaeological data. The data also suggested that the allele was introgressed into non-European *B. taurus* and *Bos indicus* breeds towards the 19th and 20th centuries, achieving an almost global distribution in the last century. Ancient DNA analyses further revealed that this mutation has been segregating in Northwestern European cattle for at least 1,000 years. Altogether, these results implicate a major role of the *PLAG1* mutation in recent changes in body size in cattle.

Keywords: Cattle; *PLAG1*; Haplotype; Pleiotropy; Stature; Reproduction

LIST OF ABBREVIATIONS

3'-UTR	3'-untranslated region
aDNA	Ancient deoxyribonucleic acid
BCF	Binary variant call format
Bd	Breadth of the distal end
Bp	Breadth of the proximal end
BT	Breadth of trochlea
BWA	Burrows-Wheeler alignment
DPA	Depth across the processus anconaeus
BLUP	Best linear unbiased predictor
CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7 gene
CHR(N)	Chromosome
CI	Confidence interval
cm	Centimeter (10^{-2} meters)
DNA	Deoxyribonucleic acid
dEBV	Deregressed estimated breeding value
EBV	Estimated breeding value
EDAR	Ectodysplasin A receptor gene
EHH	Extended haplotype homozygosity
EM	Expectation-Maximization
FAANG	Functional Annotation of Animal Genomes
F_{ST}	Fixation index
GB	Greatest breadth
GL	Greatest length
Gbp	Giga-base-pair (10^9 base-pairs)
HapAllele	Haplotype allele
HapBlock	Haplotype block
HapGenotype	Haplotype genotype
HapLibrary	Haplotype library
HapMap	Haplotype map
HD	Illumina® BovineHD BeadChip assay

HWE	Hardy-Weinberg equilibrium
IBD	Identity-by-descent or identical-by-descent
IGV	Integrative genomics viewer
kbp	Kilo-base-pair (10^3 base-pairs)
LD	Linkage disequilibrium
Mbp	Mega-base-pair (10^6 base-pairs)
miRNA	Micro ribonucleic acid
Ne_{LD}	Effective population size estimated from linkage disequilibrium
mRNA	Messenger ribonucleic acid
MOS	Moloney murine sarcoma viral oncogene homolog
PCA	Principal components analysis
PCR	Polymerase chain reaction
PLAG1	Pleomorphic adenoma gene 1
QTL	Quantitative trait locus
RNA	Ribonucleic acid
SAM	Sequence Alignment/Map
SD	Smallest breadth of diaphysis
SDO	Smallest depth of olecranon
SE	Standard error
SHAPEIT2	Segmented HAPlotype Estimation & Imputation Tool
SNP	Single nucleotide polymorphism
UMD	University of Maryland
UV	Ultraviolet
VEP	Ensembl variant effect predictor
yBP	Years before present

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 – General considerations

1. Background

The pleomorphic adenoma gene 1 (*PLAG1*) was first described as an oncogene involved in pleomorphic adenomas of salivary glands (KAS et al., 1997). To date, *PLAG1* has orthologues described in 83 species including mammals, birds, reptiles, amphibians and fish (ENSEMBL, 2017). The canonical model of *PLAG1* contains five exons, and its predominant transcript is approximately 7.3 thousand base-pairs (kbp) long (VAN DYCK et al., 2007). However, the coding sequence covers only ~1.5 kbp of the transcript and is located in exons 4 and 5 (JUMA et al., 2016). Translation of the *PLAG1* messenger RNA (mRNA) yields a nuclear protein of ~500 amino-acid residues containing zinc finger domains capable of DNA binding (KAS et al., 1998; VAN DYCK et al., 2007). Consequently, the Plag1 protein is classified as a transcription factor.

The main known target of Plag1 is the insulin-like growth factor 2 gene (*IGF2*) (VOZ et al., 2000). Plag1 most likely up-regulates *IGF2* by binding to the embryonic P3 promoter or by acting as a promoter/enhancer facilitator (JUMA et al., 2016). Although *PLAG1* is predominantly expressed during early embryonic and fetal development (KARIM et al., 2011), several tissues can express *PLAG1* during adult life, including seminiferous epithelium, Sertoli cells, ovarian follicles, prostate, heart, spleen and small intestine (QUEIMADO et al., 1999; JUMA et al., 2016). Currently, there are three putative mechanisms of regulation hypothesized for *PLAG1* and its gene products: (i) control of transcription by the bidirectional promoter shared between *PLAG1* and the coiled-coil-helix-coiled-coil-helix domain containing 7 gene (*CHCHD7*) (KARIM et al. 2011); (ii) post-translational deacetylation and SUMOylation causing activation and inhibition of Plag1, respectively (VAN DYCK et al., 2004; ZHENG; YANG 2005); (iii) and binding of microRNAs (miRNAs) to the 3'-untranslated region (3'-UTR) of the *PLAG1* mRNA, leading to repression of translation or mRNA decay (PALLASCH et al., 2009; PATZ; PALLASCH; WENDTNER, 2010; JUMA et al., 2016).

Of particular relevance here is the role of Plag1 in the stimulation of *IGF2* expression. The Insulin-like Growth Factor (IGF) or somatomedin signaling system is a complex pathway involved in growth and reproduction that is composed by ligands, cell membrane receptors and high-affinity binding proteins. The ligands include Igf1 (encoded by *IGF1*), Igf2 (enconded by *IGF2*) and insulin (encoded by *INS*). Likewise, the cell membrane receptors include the insulin receptor, Igf1r and Igf2r, which are encoded by *INSR*, *IGF1R* and *IGF2R*, respectively. Also, there are seven high-affinity binding proteins termed Igfbp 1-7 with respective genes *IGFBP* 1-7. Together, these ligands, receptors and binding proteins control the signaling cascades of the PI3K/Akt, Ras/Raf/MEK/ERK and MAPK pathways, whose activation result in cell growth, proliferation and suppressed apoptosis (ROSENZWEIG; ATREYA, 2010; KING; WONG, 2012).

Ligands Igf1 and Igf2 share amino-acid sequence homology and act upstream of the IGF signaling system by biding to the Igf1 receptor. In terms of transcription abundance, Igf1 is preferentially produced after birth in the liver, whereas Igf2 is maternally imprinted and more predominant in early embryonic and fetal development in a wide range of tissues (BERGMAN et al., 2013). Since Igf2 has a crucial role in development, reduced growth is a predicted effect of decreased expression and translation of *PLAG1* transcripts. Indeed, mice carrying a null *PLAG1* allele exhibit marked growth retardation and reduced fertility (HENSEN et al., 2004). Another important prediction is that mutations in *PLAG1* may have an effect on growth if coding or regulatory DNA sequence is involved.

Early genome-wide association studies (GWAS) for human stature confirmed the prediction above by detecting associations on the *PLAG1* chromosomal domain (GUDBJARTSSON et al., 2008; LETTRE et al., 2008; WEEDON et al., 2008; LANGO et al., 2010). A few years later, Karim et al. (2011) mapped a major quantitative trait locus (QTL) affecting stature in cattle steers descended from two *Bos taurus* breeds diverging in height, namely Jersey (short) and Holstein (tall). This QTL was narrowed down to eight positional candidate polymorphisms spanning *PLAG1*. In particular, three of these polymorphic sites affect DNA sequence implied in the regulation of *PLAG1*, namely the 3'-UTR segment and the bidirectional promoter. Apart from humans and cattle, pigs (RUBIN et al., 2012) and horses (METZGER et al., 2013)

were also found to carry genetic variation around *PLAG1* with effects on stature.

After the milestone publication by Karim et al. (2011), associations between polymorphisms around *PLAG1* and traits related to body size, growth and reproduction have been repeatedly reported in several cattle breeds (PAUSCH et al., 2011; LITTLEJOHN et al., 2012; FORTES et al., 2012, 2013a, 2013b; NISHIMURA et al., 2012; BOLORMAA et al., 2014; SAATCHI et al., 2014; FINK et al., 2017) (Figure 1). During my time as a Master of Science student between 2011 and 2013, I participated in studies involving the *Bos indicus* lineage of cattle that paralleled the work of Karim et al. (2011) and that were published a few years later (UTSUNOMIYA et al., 2013, 2014; HARTATI et al., 2015; PEREIRA et al., 2016). These studies resulted in the discovery that *PLAG1* variants also affect body size and reproduction in *B. indicus* cattle.

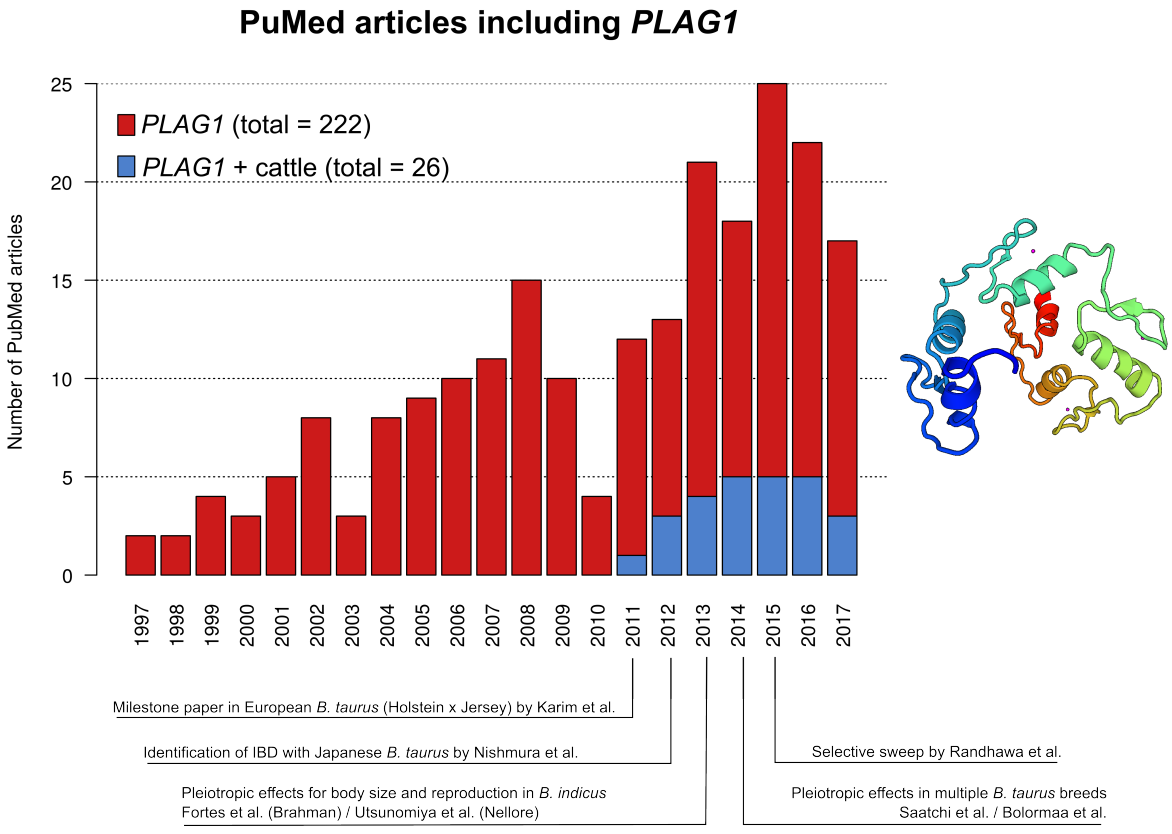

Figure 1. Published articles reporting research on *PLAG1* from 1997 to 2017.

An additional intriguing aspect of the bovine *PLAG1* chromosomal domain is that it also exhibits evidence of recent positive selection in several cattle breeds (RANDHAWA et al., 2015; BOITARD et al., 2016). Since phenotype-genotype associations and signatures of selection in the bovine *PLAG1* locus seem to be a widespread phenomenon around the world, it is tempting to hypothesize a putative role of *PLAG1* mutations in the recent evolutionary history of cattle worldwide. In particular, the main hypothesis put forth in this thesis is that the selected bovine *PLAG1* mutation contributed to recent gains in cattle stature. Settling the details of when and where the initial selective pressure might have occurred was also within the scope of the present thesis.

Although the actual mutation is still unknown, genetic inheritance is driven by segments of closely interlinked nucleotides, rather than single point mutations. Therefore, the analysis of haplotypes (i.e., chromosomal alleles formed by multiple segregating sites) was used to indirectly investigate the *PLAG1* mutation. In Chapter 2, a computer application is presented that was specifically designed and developed for this analysis. This application was published in *Bioinformatics* in June of 2016 (UTSUNOMIYA et al., 2016) as an R package (R CORE TEAM, 2017). This tool was then applied to genomic data from worldwide cattle breeds in the hope of uncovering the origin and patterns of spread of the bovine *PLAG1* mutation. Results from this investigation are reported in Chapter 3, and were published in Nature Scientific Reports in December of 2017 (UTSUNOMIYA et al., 2017).

2. References

BERGMAN, D.; HALJE, M.; NORDIN, M.; ENGSTRÖM, W. Insulin-Like Growth Factor 2 in Development and Disease: A Mini-Review. **Gerontology**, v. 59, p. 240-249, 2013.

BOITARD, S.; BOUSSAHA, M.; CAPITAN, A.; ROCHA, D.; SERVIN, B. Uncovering adaptation from sequence data: Lessons from genome resequencing of four cattle breeds. **Genetics**, v. 203, p. 433-450, 2016.

BOLORMAA, S.; PRYCE, J. E.; REVERTER, A.; ZHANG, Y.; BARENDSE, W.; KEMPER, K.; TIER, B.; SAVIN, K.; HAYES, B. J.; GODDARD, M. E. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. **PLOS Genetics**, v. 10, e1004198, 2014.

ENSEMBL. **Ensembl Genome Browser 90**. Disponível em: <http://www.ensembl.org>. Acesso em: 4 set. 2017.

FINK, T. A.; TIPLADY, K.; LOPDELL, T.; JOHNSON, T.; SNELL, R. G.; SPELMAN, R. J.; DAVIS, S. R.; LITTLEJOHN, M. D. Functional confirmation of *PLAG1* as the causative gene underlying major pleiotropic effects on liveweight and milk characteristics. **Scientific Reports**, v. 7, 44793, 2017.

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**, v. 44, p. 636-647, 2013a.

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, 2012.

FORTES, M. R. S.; REVERTER, A.; KELLY, M.; MCCULLOCH, R.; LEHNERT, S. A. Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. **Andrology**, v. 1, p. 644-650, 2013b.

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.

HARTATI, H.; UTSUNOMIYA, Y. T.; SONSTEGARD, T. S.; GARCIA, J. F.; JAKARIA, J.; MULADNO, M. Evidence of *Bos javanicus* x *Bos indicus* hybridization and major QTLs for birth weight in Indonesian Peranakan Ongole cattle. **BMC Genetics**, v. 16, 75, 2015.

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 and Differentiation**, v. 46, p. 459-470, 2004.

JUMA, A. R.; DAMDIMOPOULOU, P. E.; GROMMEN, S. V. H.; VAN DE VEN, W. J. M.; DE GROEF, B. Emerging role of *PLAG1* as a regulator of growth and reproduction. **Journal of Endocrinology**, v. 228, p. R45-R56, 2016.

KARIM, L.; TAKEDA, H.; LIN, L.; DRUET, T.; ARIAS, J. A. C.; 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, p. 405-413, 2011.

KAS, K.; VOZ, M. L.; RÖIJER, E.; ASTRÖM, A. K.; MEYEN, E.; STENMAN, G.; VAN DE VEN, W. J. Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations. **Nature Genetics**, v. 15, p. 170-174, 1997.

KAS K, VOZ ML, HENSEN K, MEYEN E & VAN DE VEN WJM. Transcriptional activation capacity of the novel PLAG family of zinc finger proteins. **Journal of Biological Chemistry**, v. 273, p. 23026-23032, 1998.

KING, E. R.; WONG, K. Insulin-like Growth Factor: Current Concepts and New Developments in Cancer Therapy. **Recent Patents on Anti-Cancer Drug Discovery**, v. 7, p. 14-30, 2012.

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.

LITTLEJOHN, M.; GRALA, T.; SANDERS, K.; WALKER, C.; WAGHORN, G.; MACDONALD, K.; COPPIETERS, W.; GEORGES, M.; SPELMAN, R.; HILLERTON, E.; DAVIS, 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, p.

591-594, 2012.

METZGER, J.; PHILIPP, U.; LOPES, M. S.; DA CAMARA MACHADO, A.; FELICETTI, M.; SILVESTRELLI, M., DISTL, O. Analysis of copy number variants by three detection algorithms and their association with body size in horses. **BMC Genomics**, v. 14, 487, 2013.

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**, v. 13, 40, 2012.

PALLASCH, C. P.; PATZ, M.; YOON, J. P.; HAGIST, S.; EGGLE, D.; CLAUS, R.; DEBEY-PASCHER, S.; SCHULZ, A.; FRENZEL, L. P.; CLAASEN, J.; KUTSCH, N.; KRAUSE, G.; MAYR, C.; ROSENWALD, A.; PLASS, C.; SCHULTZE, J. L.; HALLEK, M.; WENDTNER, C. M. miRNA deregulation by epigenetic silencing disrupts suppression of the oncogene *PLAG1* in chronic lymphocytic leukemia. **Blood**, v. 114, p. 3255-3264, 2009.

PATZ, M.; PALLASCH, C. P.; WENDTNER, C.-M. Critical role of microRNAs in chronic lymphocytic leukemia: overexpression of the oncogene *PLAG1* by deregulated miRNAs. **Leukemia & lymphoma**, v. 51, p. 1379-1381, 2010.

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.

PEREIRA, A. G. T.; UTSUNOMIYA, Y. T.; MILANESI, M.; TORRECILHA, R. B. P.; CARMO, A. S.; NEVES, H. H. R.; CARVALHEIRO, R.; AJMONE-MARSAN, P.; SONSTEGARD, T. S.; SÖLKNER, J.; CONTRERAS-CASTILLO, C. J.; GARCIA, J. F. Pleiotropic genes affecting carcass traits in *Bos indicus* (Nellore) cattle are modulators of growth. **PLOS ONE**, v. 11, e0158165, 2016.

QUEIMADO, L.; LOPES, C.; DU, F.; MARTINS, C.; BOWCOCK, A. M.; SOARES, J.; LOVETT, M. Pleomorphic adenoma gene 1 is expressed in cultured benign and malignant salivary gland tumor cells. **Laboratory Investigation**, v. 79, p. 583-589, 1999.

RANDHAWA, I. A. S.; KHATKAR, M. S.; THOMSON, P. C.; RAADSMA, H. W. Composite Selection Signals for Complex Traits Exemplified Through Bovine Stature Using Multibreed Cohorts of European and African *Bos taurus*. **Genes Genomes Genetics (G3)**, v. 5, p. 1391-1401, 2015.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.r-project.org/>. Acesso em: 4 set. 2016.

ROSENZWEIG, S. A.; ATREYA, H. S. Defining the pathway to insulin-like growth factor system targeting in cancer. **Biochemical Pharmacology**, v. 80, p. 1115-1124, 2010.

RUBIN, C. J.; MEGENS, H. J.; MARTINEZ-BARRIO, A.; MAQBOOL, K.; SAYYAB, S.; SCHWOCHOW, D.; WANG, C.; CARLBORG, Ö.; JERN, P.; JØRGENSEN, C. B.; ARCHIBALD, A. L.; FREDHOLM, M.; GROENEN, M. A.; ANDERSSON, L. Strong signatures of selection in the domestic pig genome. **Proceedings of the National Academy of Sciences of the United States of America (PNAS USA)**, v. 109, p. 19529-19536, 2012.

SAATCHI, M.; SCHNABEL, R. D.; TAYLOR, J. F.; GARRICK, D. J. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. **BMC Genomics**, v. 15, 442, 2014.

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.; 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**, v. 14, 52, 2013.

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.; SOĽKNER, J.; PORTO-NETO, L. R.; SCHENKEL, F. S.; MCEWAN, J.; COLE, J. B.; DA SILVA, M. V. G. B.; VAN TASSELL, C. P.; SONSTEGARD, T. S.; GARCIA, J. F. Genome-wide mapping of loci explaining variance in scrotal circumference in Nellore cattle. **PLOS ONE**, v. 9, e88561, 2014.

UTSUNOMIYA, Y. T.; MILANESI, M.; UTSUNOMIYA, A. T. H.; AJMONE-MARSAN, P.; GARCIA, J. F. GHap: an R package for genome-wide haplotyping. **Bioinformatics**, v. 32, p. 2861–2862, 2016.

UTSUNOMIYA, Y. T.; MILANESI, M.; UTSUNOMIYA, A. T. H.; TORRECILHA, R. B. P.; KIM, E. S.; COSTA, M. S; AGUIAR, T. S.; SCHROEDER, S.; CARMO, A. S.; CARVALHEIRO, R.; NEVES, H. H. R.; PADULA, R. C. M.; SUSSAI, T. S.; ZAVAREZ, L. B.; CIPRIANO, R. S.; CAMINHAS, M. M.; HAMBRECHT, G.; COLLI, L.; EUFEMI, E.; AJMONE-MARSAN, P.; CESANA, D.; SANNAZARO, M.; BUORA, M.; MORGANTE, M.; LIU, G.; BICKHART, D.; VAN TASSELL, C. P.; SÖLKNER, J.; SONSTEGARD, T. S.; GARCIA, J. F. A *PLAG1* mutation contributed to stature recovery in modern cattle. **Scientific Reports**, v.7, 17140, 2017.

VAN DYCK, F.; DECLERCQ, J.; BRAEM, C. V.; VAN DE VEN, W. J. M. *PLAG1*, the prototype of the PLAG gene family: Versatility in tumour development (review). **International Journal of Oncology**, v. 30, p. 765-774, 2007.

VOZ, M. L.; AGTEN, N. S.; VAN DE VEN, W. J. M.; KAS, K. *PLAG1*, the main translocation target in pleomorphic adenoma of the salivary glands, is a positive regulator of IGF-II. **Cancer Research**, v. 60, p. 106-113, 2000.

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.

ZHENG, G.; YANG, Y. C. SUMOylation and acetylation play opposite roles in the transactivation of PLAG1 and PLAGL2. **Journal of Biological Chemistry**, v. 280, p. 40773-40781, 2005.

CHAPTER 2 - GHap: An R package for Genome-wide Haplotyping

Utsunomiya, Y.T.; Milanesi, M.; Utsunomiya, A.T.H.; Ajmone-Marsan, P.; Garcia, J.F.

Bioinformatics 32(18):2861-2862, 2016
DOI: 10.1093/bioinformatics/btw356

1. Abstract

The GHap R package was designed to call haplotypes from phased marker data. Given user-defined haplotype blocks (HapBlock), the package identifies the different haplotype alleles (HapAllele) present in the data and scores sample haplotype allele genotypes (HapGenotype) based on HapAllele dose (i.e., 0, 1 or 2 copies). The output is not only useful for analyses that can handle multi-allelic markers, but is also conveniently formatted for existing pipelines intended for bi-allelic markers.

Keywords: Multi-marker analysis; Genetic variant; SNP; software

2. Introduction

The use of high-density marker panels in genomics relies on the concept of linkage disequilibrium (LD) and tagging, such that information from unobserved variants can be indirectly captured by correlation with nearby markers (BUSH; MOORE, 2012). Methods for genomic analysis are usually based on single markers, ignoring that unobserved variants may be better modeled by the use of phase information (BROWNING; BROWNING, 2008). Moreover, the need for tools to perform haplotype calls from phased data has been under-served in spite of the growing interest in haplotype-based analyses in the last years. Here we describe GHap, an R package designed to call haplotypes from phased data. The goal of GHap is to compute summary statistics for haplotype blocks (HapBlock) and haplotype alleles (HapAllele), as well as to construct a matrix of haplotype genotypes (HapGenotype). As a general framework, each HapAllele can be treated as a pseudo-marker in down-stream analyses, facilitating the incorporation of phase information in existing pipelines. This approach differs from competing methods as it uses HapAllele as markers, instead of hidden haplotype states generated by expectation maximization or hidden Markov models.

3. Implementation

3.1. Loading and manipulating data

The GHap input format is described in APPENDIX A and can be derived from popular phasing programs such as SHAPEIT2 (O'CONNELL et al., 2014). GHap assumes that family information, if applicable, has been taken into account during phasing. As GHap assumes known phase, low quality phasing will directly impact the reliability of haplotype calls. The phased data is loaded using *ghap.loadphase()*. The *ghap.maf()* function can be used to identify markers with low polymorphic information content. The *ghap.subsetphase()* function subsets the data by inactivating specified samples and markers, while *ghap.mergephase()* combines different phased data. The data can be exported using *ghap.outphase()*.

3.2. Genome-wide haplotyping procedure

Let a HapBlock be a user-defined set of adjacent markers and the haplotype library (HapLibrary) be the collection of observed HapAlleles for that HapBlock. The haplotyping procedure implemented in the *ghap.haplotyping()* function is straightforward: each HapAllele in the HapLibrary is treated as a marker, and HapGenotypes are scored as 0, 1 or 2 allele copies (for more information see APPENDIX A). HapGenotypes can then be loaded into R using *ghap.loadhaplo()* and manipulated with *ghap.subsethaplo()* and *ghap.mergehaplo()*, or exported to text file and the transposed format from PLINK (PURCELL et al., 2007) with *ghap.outhaplo()* and *ghap.hap2tped()*, respectively. Typically, the user may want to target a specific HapBlock based on prior information (e.g. pre-computed LD blocks). We also provide an alternative approach through the *ghap.blockgen()* function, which allows the user to specify arbitrary windows and step size based on markers or segments.

3.3. Haplotype statistics and auxiliary functions

The *ghap.hapstats()* function computes a series of summary statistics for HapAlleles, which includes: number of observations, frequency, observed number of homozygotes and heterozygotes, expected number of homozygotes and three different measures of deviations from Hardy-Weinberg equilibrium. We also implemented a series of auxiliary functions, namely *ghap.blockstats()*, *ghap.fst()*, *ghap.ancestral()*, *ghap.kinship()*, *ghap.pca()*, *ghap.blmm()*, *ghap.assoc()*, *ghap.blup()* and *ghap.simpheno*, to estimate block expected heterozygosity and number of HapAlleles, haplotype-based F_{ST}, HapAllele origin, relationship matrices, principal components, linear mixed models, association analyses, Best Linear Unbiased Predictor (BLUP) and simulate phenotypes, respectively. Additionally, given a vector of arbitrary scores for HapAlleles, the *ghap.profile()* function allows for computing individual profiles as sum(HapGenotypes*scores).

4. Examples

We tested GHap using reference phased data available at the IMPUTE2 HOWIE; DONELLY; MARCHINI, 2009) software website (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference). These data derive from the HapMap Project Phase 3 (THE INTERNATIONAL HAPMAP 3 CONSORTIUM, 2010), and comprise 1011 subjects from 11 human populations and 20,000 random SNPs mapping to chromosome 2. This dataset is available through the *ghap.makefile()* function. Benchmarking of the main tasks showed that elapse time scaled linearly with increasing number of samples and markers in a trial up to 100,000 SNPs and 50,000 subjects (APPENDIX A). We performed three analyses: (i) Principal Components Analysis (PCA); (ii) detection of divergent loci between Chinese and Europeans and (iii) mixed model association analysis with phenotypes simulated based on the real genotypes. Similar analyses can be done following the package documentation. Although based on a small example set of markers in a single chromosome, the haplotype calls generated by GHap resolved the known genetic structure in the HapMap dataset (Figure 1A). The haplotype-based F_{ST} analysis (Figure 1B) identified a previously reported signature of selection in Chinese encompassing *EDAR* (SABETI et al., 2007), with the top scoring HapBlock mapping to its intragenic region. Finally, the association analysis (Figure 1C) efficiently pinpointed the HapAllele segregating with the simulated causal variant.

5. Conclusion

The GHap package provides means for haplotype calling, facilitating the incorporation of phase information in genome-wide analyses. The package is available at: https://cran.r-project.org/package=GHap.

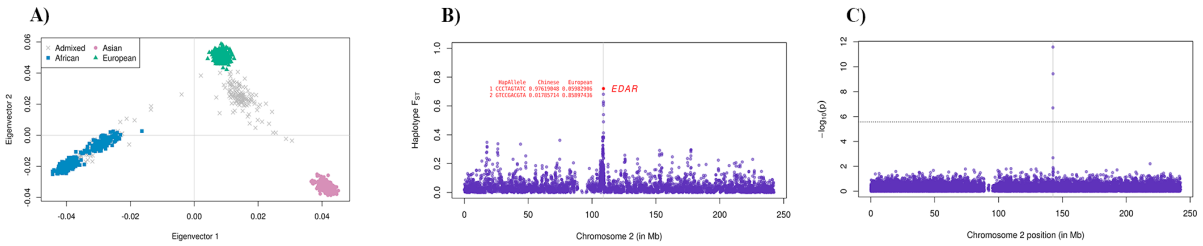

Figure 1. Examples of applications of the GHap package with Human HapMap Project Phase 3 data. Analyses were based on HapBlocks of 10 markers with overlaps of 5 markers between consecutive blocks. (A) Clustering of subjects based on a PCA of the HapAllele relationship matrix. (B) Haplotype-based F_{ST} for Chinese x European. (C) Mixed model association analysis using simulated phenotypes. The vertical grey line marks the position of the simulated causal nucleotide, and the horizontal dashed line marks the Bonferroni significance threshold ($p < 2.73 \times 10^{-6}$)

6. Funding

This research received financial support from: Sao Paulo Research Foundation (FAPESP - http://www.fapesp.br/) (process 2014/01095-8); The National Council for Scientific and Technological Development (CNPq - http://www.cnpq.br/) (process 407502/2013-0). Conflict of Interest: none declared.

7. References

BROWNING, B.L.; BROWNING,S.R. Haplotypic analysis of Wellcome Trust Case Control Consortium data. **Human Genetics**, v. 123, p. 273-280, 2008.

BUSH,W.S.; MOORE, J. H. Chapter 11: Genome-wide association studies. **PLOS Computational Biology**, v. 8, e1002822, 2012.

HOWIE, B.N.; DONELLY, P.; MARCHINI, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.

PLOS Genetics, v. 5, e1000529, 2009.

O'CONNELL, J.; GURDASANI, D.; DELANEAU, O.; PIRASTU, N.; ULIVI, S.; COCCA, M.; TRAGLIA, M.; HUANG, J.; HUFFMAN, J. E.; RUDAN, I.; MCQUILLAN, R.; FRASER, R. M.; CAMPBELL, H.; POLASEK, O.; ASIKI, G.; EKORU, K.; HAYWARD, C.; WRIGHT, A. F.; VITART, V.; NAVARRO, P.; ZAGURY, J. F.; WILSON, J. F.; TONIOLO, D.; GASPARINI, P.; SORANZO, N.; SANDHU, M. S.; MARCHINI, J. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. **PLOS Genetics**, v. 10, e1004234, 2014.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I. W.; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, v. 81, p. 559-575, 2007.

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.; THE INTERNATIONAL HAPMAP CONSORTIUM. Genome-wide detection and characterization of positive selection in human populations. **Nature**, v. 449, p.913-918, 2007.

THE INTERNATIONAL HAPMAP 3 CONSORTIUM. Integrating common and rare genetic variation in diverse human populations. **Nature**, v. 467, 52-58, 2010.

CHAPTER 3 - A *PLAG1* mutation contributed to stature recovery in modern cattle

Utsunomiya, Y.T.; Milanesi, M.; Utsunomiya, A.T.H.; Torrecilha, R.B.P.; Kim, E.S.; Costa, M.S.; Aguiar, T.S.; Schroeder, S.; Carmo, A.S.; Carvalheiro, R.; Neves, H.H.R.; Padula, R.C.M.; Sussai, T.S.; Zavarez, L.B.; Cipriano, R.S.; Caminhas, M.M.T.; Hambrecht, G.; Colli, L.; Eufemi, E.; Ajmone-Marsan, P.; Cesana, D.; Sannazaro, M.; Buora, M.; Morgante, M.; Liu, G.; Bickhart, D.; Van Tassell, C.P.; Sölkner, J.; Sonstegard, T.S.; Garcia, J.F.

Scientific Reports 7(1):17140, 2017
DOI: 10.1038/s41598-017-17127-1

1. Abstract

The recent evolution of cattle is marked by fluctuations in body size. Height in the *Bos taurus* lineage was reduced by a factor of ~1.5 from the Neolithic to the Middle Ages, and increased again only during the Early Modern Ages. Using haplotype analysis, we found evidence that the bovine *PLAG1* mutation (*Q*) with major effects on body size, weight and reproduction is a >1,000 years old derived allele that increased rapidly in frequency in Northwestern European *B. taurus* between the 16th and 18th centuries. Towards the 19th and 20th centuries, *Q* was introgressed into non-European *B. taurus* and *Bos indicus* breeds. These data implicate a major role of *Q* in recent changes in body size in modern cattle, and represent one of the first examples of a genomic sweep in livestock that was driven by selection on a complex trait.

Keywords: Bovine; SNP; Height; Pleomorphic Adenoma Gene 1

2. Introduction

The extinct wild auroch (*Bos primigenius*) lost stature during late Pleistocene, decreasing from a withers height range of 165 – 185 cm to 145 – 160 cm (GUINTARD, 1999). Between 280,000 (MURRAY et al., 2010) and 330,000 (ACHILLI et al., 2008) years before present (yBP), the ancestral auroch population diverged into two distinct lineages that would later originate the humpless *Bos taurus* and the humped *Bos indicus* cattle. Towards the beginning of the Holocene, *B. taurus* and *B. indicus* were independently domesticated in the Fertile Crescent (~10,500 yBP) and in the Indus Valley (~8,500 yBP), respectively (LOFTUS et al., 1994; BRUFORD; BRADLEY; LUIKART, 2003). Later, *B. taurus* suffered a further decline in stature between the Neolithic and the Early Middle Ages (AJMONE-MARSAN; GARCIA; LENSTRA, 2010), approaching wither sizes of 95 – 123 cm (GUINTARD, 1999). Archaeological data further suggested that stature loss followed a gradient from Southwestern Asia towards Northwestern Europe that coincides with the post-domestication route of Europe colonization (LASOTA-MOSKALEWSKA; KOBRYN, 1990). A counterpoint to this process was also suggested, namely introduction of cattle in Northern and Western Europe by the Roman Empire around the 1st century that were much larger than Celtic or Germanic cattle, ranging from 105 to 142 cm (GUINTARD, 1999). However, stature of Northwestern cattle decreased again shortly after the fall of the Roman Empire. Along with the intensification of artificial selection in the past few centuries, *B. taurus* entered a process of stature recovery that started between the 15th and 18th centuries and that lasted until recently (GUINTARD 1999, AJMONE-MARSAN; GARCIA; LENSTRA, 2010). Consequently, modern European breeds typically range from 105 to 155 cm in average withers height (FAO, 2017), which represents a 1.10- to 1.26-fold increase in stature compared to the Early Middle Ages (Figure 1).

If major genetic variants contributed to stature recovery since the Early Modern Ages, their selection signatures should be detectable from genomic data of modern cattle breeds. Randhawa et al. (2015) contrasted genome-wide single nucleotide polymorphism (SNP) data of European *B. taurus* breeds with high (145 – 155 cm) and low (105 – 133 cm) median withers height and found that, relative to the

UMD v3.1 genome assembly (ZIMIN et al., 2009), the most significant signature related to stature mapped to chromosome 14 (CHR14) positions 24.79 – 28.25 million base-pairs (Mbp). This signature has been recently confirmed with whole genome sequence data of four *B. taurus* breeds and constrained to a smaller region spanning positions 24.80 – 25.08 Mbp (BOITARD et al., 2016), where the pleomorphic adenoma gene 1 (*PLAG1*) is located. Nevertheless, it is still unclear whether this selective sweep is traceable back to the period between the 15th and 18th centuries.

Following the milestone publication by Karim et al. (2011), we and others (FORTES et al., 2012, 2013a, 2013b; LITTLEJOHN et al., 2012; NISHIMURA et al., 2012; UTSUNOMIYA et al., 2013, 2014b; SAATCHI et al., 2014; HARTATI et al., 2015; PEREIRA et al., 2016) have previously reported that this CHR14 region is also a major pleiotropic quantitative trait locus (QTL) affecting cattle body size and reproduction, supporting an important role of *PLAG1* in recent changes in stature in cattle. The candidacy of *PLAG1* is further supported by functional evidence, since the transcription factor encoded by this gene regulates the expression of insulin-like growth factors (QUEIMADO et al., 1999; VOZ et al., 2000; VAN DYCK et al., 2007; FORTES et al., 2013a; JUMA et al., 2016). Moreover, mice carrying a null *PLAG1* allele exhibit growth retardation and reduced fertility (HENSEN et al., 2004). An intriguing observation however is that the QTL is detectable in modern breeds of both *B. taurus* and *B. indicus* cattle, which is unexpected given the long divergence time between these two subspecies.

Three main hypotheses are consistent with a major stature QTL that is detectable in both *B. taurus* and *B. indicus*: (i) variants that were present in the ancestral *B. primigenius* population still segregate in both subspecies; (ii) separate derived alleles account for the QTL in the two subspecies; or (iii) lineage-specific mutations have been recently introgressed from one subspecies to the other. Together with the evidence of selection, hypothesis (i) would imply that very ancient (> 280,000 years old) standing neutral variants at the *PLAG1* locus became differentially advantageous under certain circumstances. This hypothesis could be confirmed by showing that ancestral haplotypes have positive effects on body size in both subspecies. On the other hand, hypotheses (ii) and (iii) would require derived

haplotypes, and could be tested by showing that *B. taurus* and *B. indicus* either carry different (ii) or identical (iii) derived haplotypes positively increasing body size.

Here, we aimed at elucidating the patterns of gene flow of *PLAG1* haplotypes by gathering evidence from the *B. indicus* lineage and contrasting against data of worldwide *B. taurus* breeds. It was our objective to investigate whether the selective sweep surrounding the haplotype was traceable to the recent event of stature recovery in Northwestern Europe.

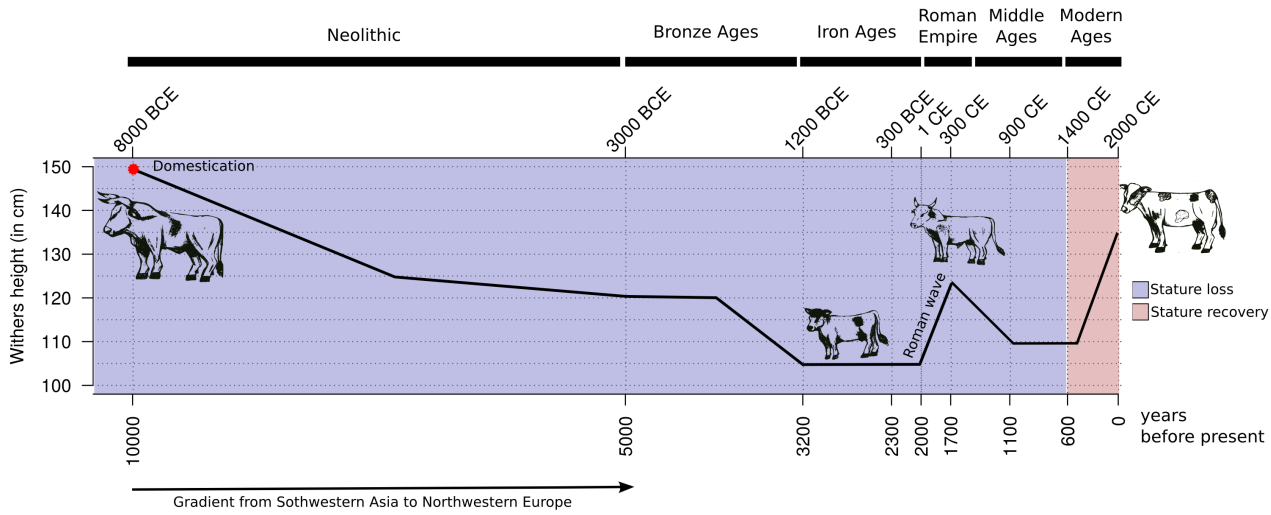

Figure 1. Schematic of clinal and temporal variation in cattle stature. The line art was produced in R v.3.3.2 (R CORE TEAM, 2017) and enhanced in Inkscape v0.48.4-r9939 (THE INKSCAPE TEAM, 2017).

3. Results

3.1. Association analysis in *B. indicus* maps a derived haplotype tagging the *PLAG1* mutation (*Q*)

In support to the hypothesis of introgression, Fortes et al. (2013a) presented evidence that the QTL haplotype associated with increased body size has an exclusive *B. taurus* origin in Brahman, a breed with average 91% *B. indicus* and 9% *B. taurus* ancestry (PÉREZ-O'BRIEN et al., 2015). Although Brahman could be an exception given its higher levels of *B. taurus* ancestry in comparison to breeds that

are generally considered purebred *B. indicus*, similar heterogeneity of haplotype origin at this QTL in other breeds is plausible since most of the modern *B. indicus* populations were recently subjected to *B. taurus* introgression (MCTAVISH et al., 2013; DECKER et al., 2014; UTSUNOMIYA et al., 2014a).

In order to test whether the introgression hypothesis holds in other *B. indicus* populations, we decided to re-map the QTL in Brazilian Nellore cattle. Our past investigation in this breed (UTSUNOMIYA et al., 2013, 2014b; PEREIRA et al., 2016) pointed to a haplotype with positive effects on weight and conformation traits, in agreement with the findings by Fortes et al. (2013a). However, as for other *B. indicus* breeds, a *B. taurus* origin for this haplotype in Nellore is less clear since we found that the current population has on average only 0.9% *B. taurus* ancestry (UTSUNOMIYA et al., 2014a; PÉREZ-O'BRIEN et al., 2015). Nevertheless, the introgression hypothesis remains plausible given the documented crossbreeding in Brazil between *B. taurus* brought by colonizers and *B*. *indicus* imported from India in the late 19th century and early 20th century (AJMONE-MARSAN; GARCIA; LENSTRA, 2010). Moreover, considering a genome size of ~2.87 billion base-pairs (Gbp) with ~22,000 protein coding genes (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM et al., 2009), 0.9% would correspond to ~25.8 Mbp of sequence, which is approximately one fourth of an average-sized chromosome (~100 Mbp) or ~200 protein-coding genes.

We started by analyzing Illumina® BovineHD (HD) genotypes of a sample of 779 Nellore bulls with deregressed estimated breeding values (dEBVs) for birth weight obtained from records of 846,782 calves. These bulls sired over ten Nellore generations and comprised animals born between 1965 and 2008, with 90% born after 1990 and 50% after 2000. We selected birth weight as a proxy for body size because: (i) withers height measurements were not available for these animals; (ii) the QTL was associated with differences in fetal expression of *PLAG1* in *B. taurus* (KARIM et al., 2011); and (iii) birth weight is moderately heritable (h^2 = 0.37) and affected by fewer environmental effects in comparison to other massively recorded traits in Nellore cattle. Whole chromosome haplotypes were constructed from 15,132 SNP markers on CHR14 with a Hidden Markov Model (O'CONNELL et al., 2014). Haplotypes at short segments were then determined in order to map the QTL via

regression analysis.

We were able to map associations to a ~39.5 thousand base-pairs (kbp) haplotype (p = 4.83 x 10^{-17}) spanning positions CHR14:24973324-25012733 (Figure 2a). Importantly, this segment was a subset of the ~271 kbp region reported by Boitard et al. (2016) in their selection signature analysis using *B. taurus* sequence data. This region contained the v-mos Moloney murine sarcoma viral oncogene homolog (*MOS*) and a portion of the 3' end of *PLAG1*. The significant haplotype included nucleotides G – rs110243083, G – rs136888475, G – rs109636480, T – rs135404594, T – rs134286310 and C – rs135538206, and will be hereafter denoted GGGTTC. Marker rs135404594 was located in the 3'-UTR of *PLAG1*, whereas markers rs134286310 and rs135538206 were intronic to *PLAG1*. All remaining markers in the haplotype were intergenic. Frequency of GGGTTC was 17.8%, with an estimated effect of 0.311 ± 0.037 kg on birth weight dEBVs. The distribution of dEBVs according to number of copies of GGGTTC followed an additive pattern (Figure 2b), confirming previous reports (KARIM et al., 2011; UTSUNOMIYA et al., 2013).

For the sake of simplicity, we will denote *Q* the unknown causal allele with positive effect on birth weight, whereas *q* will refer to the alternative allele, following notation introduced by Karim et al. (2011). Assuming *Q* co-segregates with the GGGTTC tag in Nellore, 18 (2.3%), 241 (30.9%) and 520 (66.8%) bulls had predicted genotypes *QQ*, *Qq* and *qq*, respectively, which did not significantly deviate from Hardy-Weinberg Equilibrium (HWE, p = 0.215). Among the six alternative haplotypes, the most common was ATACCT (70.4%). An orthology analysis revealed ATACCT to be the ancestral haplotype (Figure 2c). Therefore, *q* and *Q* were most likely the ancestral and derived alleles, respectively, in agreement with Fortes et al. (2013a). Consequently, the data provided little support for the hypothesis of *Q* being a mutation already present in the population of aurochsen ancestral to both *B. taurus* and *B. indicus* cattle. Furthermore, the existence of other low frequency haplotypes in this sample suggested that *q* occurs in several different haplotype backgrounds, consistent with an ancestral allele.

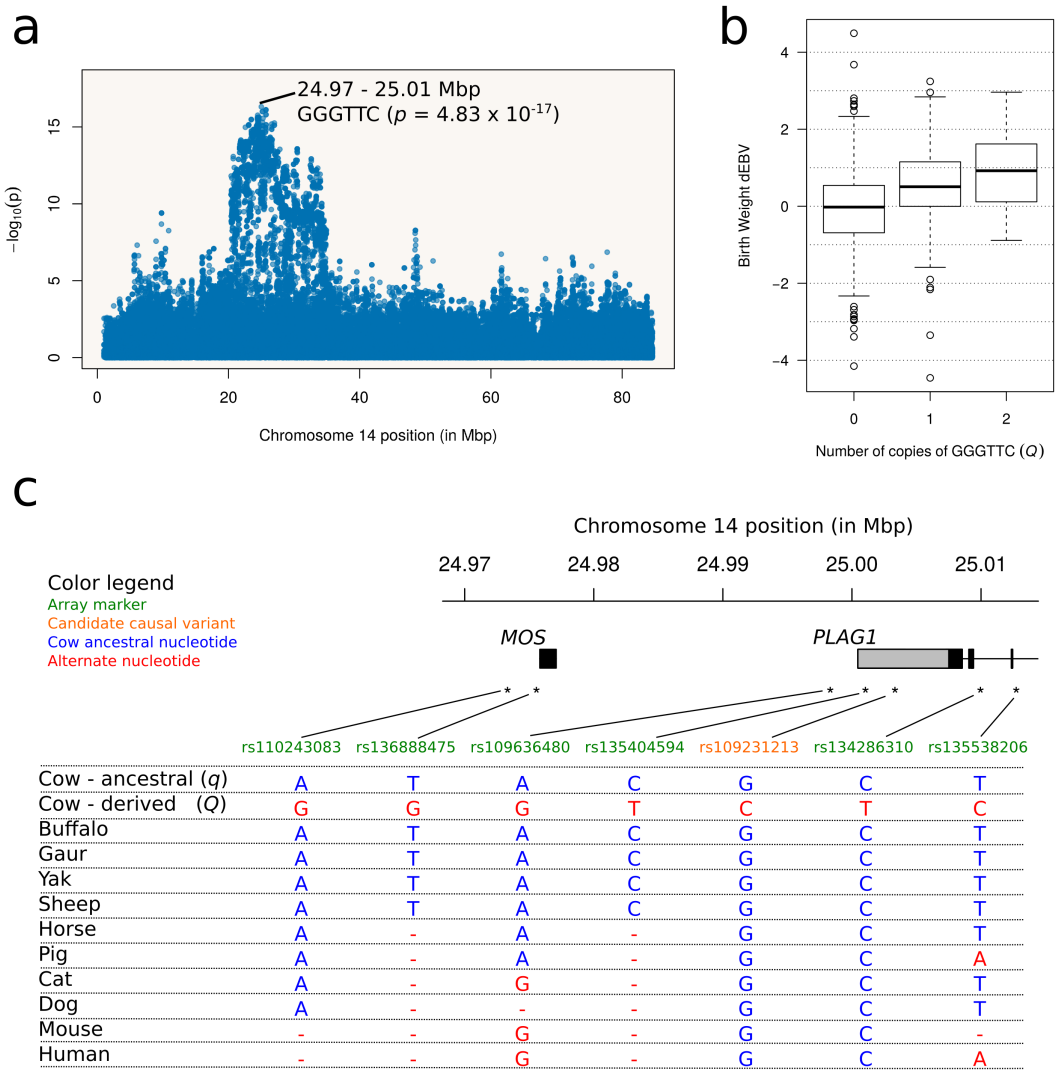

Figure 2. Identification of a haplotype tagging the *PLAG1* mutation (*Q*) in *B. indicus*. (a) Scatterplot showing the birth weight (dEBV) haplotype association mapping on chromosome 14 in Nellore cattle. Maximum association (p = 4.83 x 10^{-17}) was detected in a ~39.5 kbp segment spanning positions 24973324 to 25012733, where *MOS* and the 3' end of *PLAG1* are located. (b) The distribution of birth weight dEBVs according to number of copies of the tag haplotype indicates an additive effect of 0.311 ± 0.037 kg. (c) Orthology analysis suggesting *Q* to be a derived mutation.

3.2. Ancestry and sequence analysis in *B. indicus* reveals *B. taurus* introgression

In order to test the hypothesis of two separate derived mutations in *B. taurus* and *B. indicus*, we searched for carriers of *Q* among Nellore bulls that directly descended from animals imported from India in the 20th century and, therefore, were unlikely to carry *B. taurus* ancestry. Our sample contained 15 animals meeting this criterion. Given that the variant was under HWE, four to five heterozygous bulls were expected assuming a heterozygosity range of 27.6 – 34.2% (95% CI considering a standard error of 1.7%) but all 15 were predicted to be *qq*. In fact, the earliest predicted carrier of *Q* was a polled animal born in 1975. Coincidently, the polled allele in Brazilian Nellore (UTSUNOMIYA et al., 2016a) is deemed to be of *B. taurus* origin. We further performed a model-based clustering analysis (ALEXANDER; NOVEMBRE; LANGE, 2009) to estimate *B. taurus* ancestry on CHR14 using a panel of reference breeds (THE BOVINE HAPMAP CONSORTIUM, 2009; PORTO-NETO et al., 2013). Estimated percentages ranged from 0.0% to 46.9% (mean = 10.0 ± 9.55%) and were positively associated with birth weight dEBVs (p = 3.31 x 10^{-9}) and number of copies of GGGTTC (r = 0.845, p = 3.60 x 10^{-213}), indicating a *B. taurus* origin for *Q* in the *B. indicus* lineage (Figure 3a). Refinement of ancestry estimates at the level of individual loci revealed that the recombination break points of introgressed segments mimicked the topology of the haplotype-based association analysis, suggesting that birth weight associations on this chromosome were essentially driven by haplotypes of *B. taurus* origin. The coordinate CHR14:24459302-25246448 exhibited the highest *B. taurus* introgression, which was estimated at 18.64% across individuals. Remarkably, this ~787.2 kbp segment contained the ~39.5 kbp region identified earlier through association analysis, and its ancestry estimate was very close to the frequency of the GGGTTC haplotype.

As genetic stratification into two breeding subgroups has been reported in the Nellore breed (UTSUNOMIYA et al., 2013; NEVES et al., 2014), we decided to investigate whether *B. taurus* ancestry on CHR14, frequency of *Q*, mean birth weight dEBV and frequency of polledness further differed between these subgroups. Subgroup 1 was known to be selected for production traits, whereas subgroup 2 has

been selected mainly for breed type. Also, breeders of subgroup 1 aim at average birth weight in order to maintain positive genetic trends for calving ease. We were able to retrieve horned/polled phenotypes for a subset of 379 bulls and assign individuals to breeding subgroups via a k-means clustering algorithm (HARTIGAN; WONG, 1979). In comparison with subgroup 1, subgroup 2 presented larger average *B. taurus* ancestry on CHR14 (10.5% against 6.2%, $p = 3.37 \times 10^{-4}$), higher frequency of *Q* (20.1% against 5.7%, $p = 2.08 \times 10^{-8}$), higher incidence of polledness (14.4% against 0.0%, $p = 0.03$), and larger mean birth weight dEBV (0.199 kg against -0.090 kg, $p = 9.76 \times 10^{-3}$) (Figure 3b). These results further supported the positive effect of *Q* on weight and a *B. taurus* origin for both *Q* and polledness in Nellore cattle.

Next, we sequenced the whole genomes of 24 Nellore bulls at ~9x coverage. Nine of these bulls were predicted to be *Qq* and fifteen were predicted to be *qq* based on number of copies of GGGTTC. These data were then compared against sequence variants underlying the stature QTL in *B. taurus*: by analyzing crossbred steers descended from two *B. taurus* dairy breeds diverging in height (i.e., small Jerseys and large Holsteins), Karim et al. (2011) narrowed the QTL down to eight positional candidate variants (APPENDIX B). Nishimura et al. (2012) also found these variants in Japanese Black cattle, and reported that the direction of the haplotype effect was consistent between studies. These findings implied *Q* to be identical-by-descent across *B. taurus* breeds. An additional implication from the introgression hypothesis was that a single derived haplotype should account for QTL effects in both subspecies of cattle. Here, the coordinates of the associated haplotype region spanned five out of the eight *B. taurus* positional candidates, namely rs110092040 (CHR14:24973953, T>C), ss319607399 (CHR14:24974221, A>G, also rs208989386), ss319607400 (CHR14:24974811, A>G, also rs137303549), rs109231213 (CHR14:25003338, C>G) and ss319607401 (CHR14:25006125, T>C, also rs134215421). The previously reported *B. taurus* haplotype at these variants was *Q* = TAACT. We found that the HD haplotype GGGTTC was in perfect correspondence with the sequence haplotype TAACT in Nellore, suggesting near perfect LD between GGGTTC and *Q* in this breed. Moreover, given that the reference genome animal was most likely *QQ* based on its GGGTTC haplotype, predictions of *qq* genotypes were corroborated by deficit of reference alleles (Figure 3c).

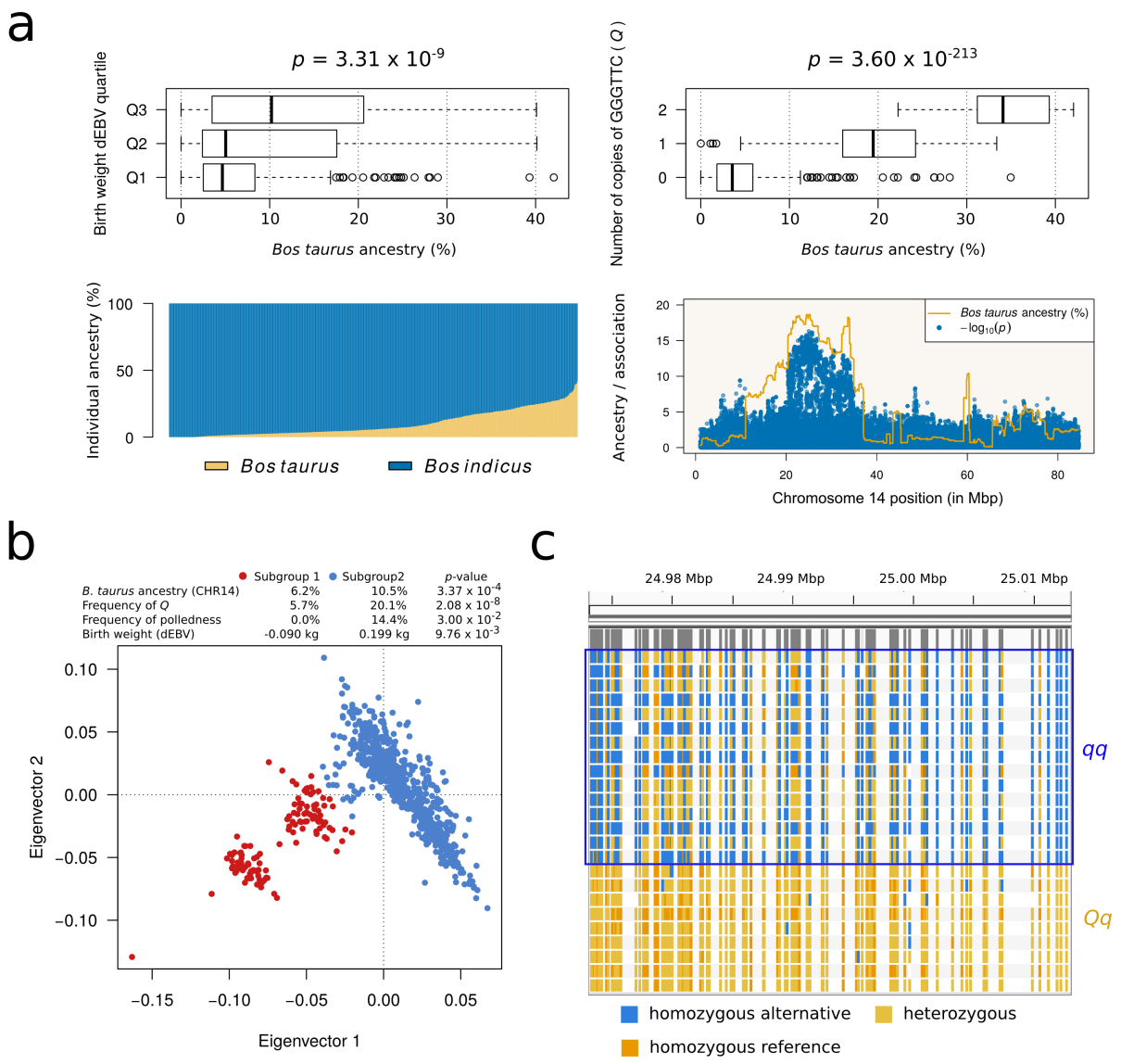

Figure 3. *B. taurus* introgression as a source for *Q* in *B. indicus*. (a) Ancestry analysis showing that both birth weight dEBVs and occurrence of *Q* were positively associated with percentage of *B. taurus* ancestry on chromosome 14. (b) A principal components analysis revealed two breeding subgroups of Nellore cattle differing in *B. taurus* ancestry, frequency of *Q*, incidence of polledness (presumably also resulting from *B. taurus* introgression) and mean of birth weight dEBVs. (c) Next-generation sequencing data of 24 Nellore bulls confirming a *B. taurus* origin of *Q* in *B. indicus*. Each row represents a bull, and colored vertical bars represent genotypes at different variant sites.

3.3. Haplotype diversity in worldwide cattle indicates selection for *Q* in Northwestern Europe

Given the *Q* allele has been recently transferred from *B. taurus* to *B. indicus*, GGGTTC should be the selected haplotype in European breeds and its frequency should be correlated with average body weight and size. In order to test this prediction, we explored haplotype frequencies in the Bovine HapMap data (THE BOVINE HAPMAP CONSORTIUM, 2009; PORTO-NETO et al., 2013), which included HD genotypes from 20 *B. taurus* (n = 503) and two *B. indicus* (n = 65) breeds. Additionally, samples of five *B. taurus* x *B. indicus* crossbred populations (n = 139) and three outgroup species (n = 11) were also available.

The GGGTTC tag was found to be the predominant haplotype in *B. taurus* breeds (APPENDIX B), consistent with the previously reported selective sweep RANDHAWA et al., 2015; BOITARD et al., 2016). However, the moderate frequency of GGGTTC in breeds that a *priori* segregate the *Q* mutation at low frequency suggested that LD between GGGTTC and *Q* may be imperfect in *B. taurus* cattle, as opposed to a near perfect LD found in the Nellore population. For instance, Karim et al. (2011) estimated that Jersey cattle carry *Q* at a frequency of ~5%, a much lower frequency than 35.9% observed for the GGGTTC haplotype here. Indeed, the authors also reported that the *q* chromosome of one out of four *Qq* sires in their data shared an identical-by-state haplotype with the *Q* chromosome when only SNP array markers were considered. These findings suggest that *Q* frequency predicted from the number of copies of GGGTTC could be overestimated in *B. taurus* as this tag haplotype may also segregate with the ancestral *q* allele at a lower frequency.

In order to address the issue of imperfect LD between GGGTTC and *Q* in *B. taurus*, we took advantage of a variant that is present in the HD array and that is in near perfect LD with the causal mutation in the Jersey and Holstein breeds (Karim et al., 2012). This variant, namely rs109815800 (CHR14:25015640, T>G), is positioned immediately after GGGTTC in the HD panel and is located only ~2.9 kbp downstream of the tag haplotype region. Therefore, we attempted to characterize LD between GGGTTC and *Q* by computing the correlation between the tag haplotype and the nucleotide G at rs109815800 in the two mentioned breeds. Our analysis indicated a

correlation of r = 0.884 (p = 1.02 x 10^{-39}), reinforcing the hypothesis that GGGTTC-*Q* occurs at a higher frequency than GGGTTC-*q* in *B. taurus*. Extrapolating from Holstein and Jersey and assuming that the gametic phase of *Q* and G-rs109815800 is at least partially preserved across cattle breeds, we found r = 0.889 (p = 3.98 x 10^{-245}) when all breeds were analyzed simultaneously. We further inspected how often G and T at rs109815800 segregated with GGGTTC in the overall data and found that G-rs109815800 happened almost exclusively with GGGTTC while accounting for 86.9% of all instances of the tag haplotype. Therefore, LD between GGGTTC and *Q* was indeed imperfect but yet substantially high across *B. taurus* populations, which validates this tag haplotype as a proxy for the causal mutation in the present study. However, in order to maximize power, we decided to extend GGGTTC to include G-rs109815800 (GGGTTCG hereafter) for a more robust assessment of the distribution of *Q* in worldwide cattle.

Frequency of GGGTTCG ranged from 13.0% – 100.0% in Northwestern European cattle (Figure 4). On the other hand, GGGTTCG was less common in Central and Southern European breeds (0.0% – 20.0%) and even rarer in *B. indicus* (0.0% – 4.0%). Also, GGGTTCG was absent in African *B. taurus*. These data indicated a Northwestern European origin for *Q* or a higher intensity of selection for body weight and size in Northwestern Europe. Haplotype ATACCTT was the most common among *B. indicus* and outgroup populations, confirming its status as ancestral haplotype and supporting further the *B. taurus* origin of *Q*. Additionally, frequency of GGGTTCG in *B. taurus* was positively correlated with average body weight (r = 0.607, p = 9.99 x 10^{-4}) and average withers height (r = 0.326, p = 0.120), and GGGTTCG was nearly fixed in admixed and crossbred populations selected for body size and weight (APPENDIX B). Geographical interpolation of the frequency of GGGTTCG further suggested an area facing the Atlantic facade of France, Belgium, Netherlands, Germany, Denmark, Norway, Sweden and Britain as the most likely centre of selection for Q (Figure 5).

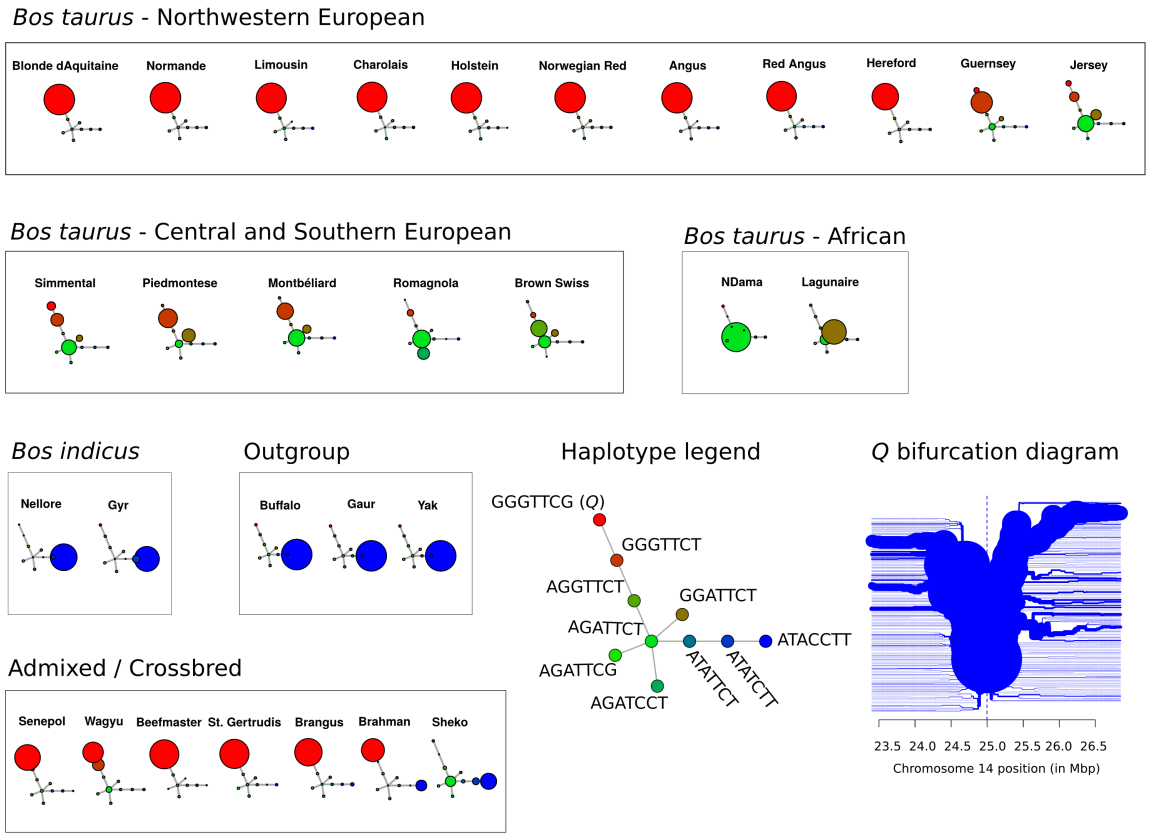

Figure 4. Haplotype diversity at the *PLAG1* locus in the Bovine HapMap data (THE BOVINE HAPMAP CONSORTIUM, 2009; PORTO-NETO et al., 2013). Each node represents a haplotype and edges connect nodes sequentially differing in one or two nucleotides. Node size is proportional to haplotype frequency. The *Q*-tagging haplotype is shown to be highly frequent in breeds originated from Northwestern Europe. A bifurcation diagram (rooted at rs109815800) is also shown, portraying the long-range linkage disequilibrium (LD) and low haplotype diversity around *Q*.

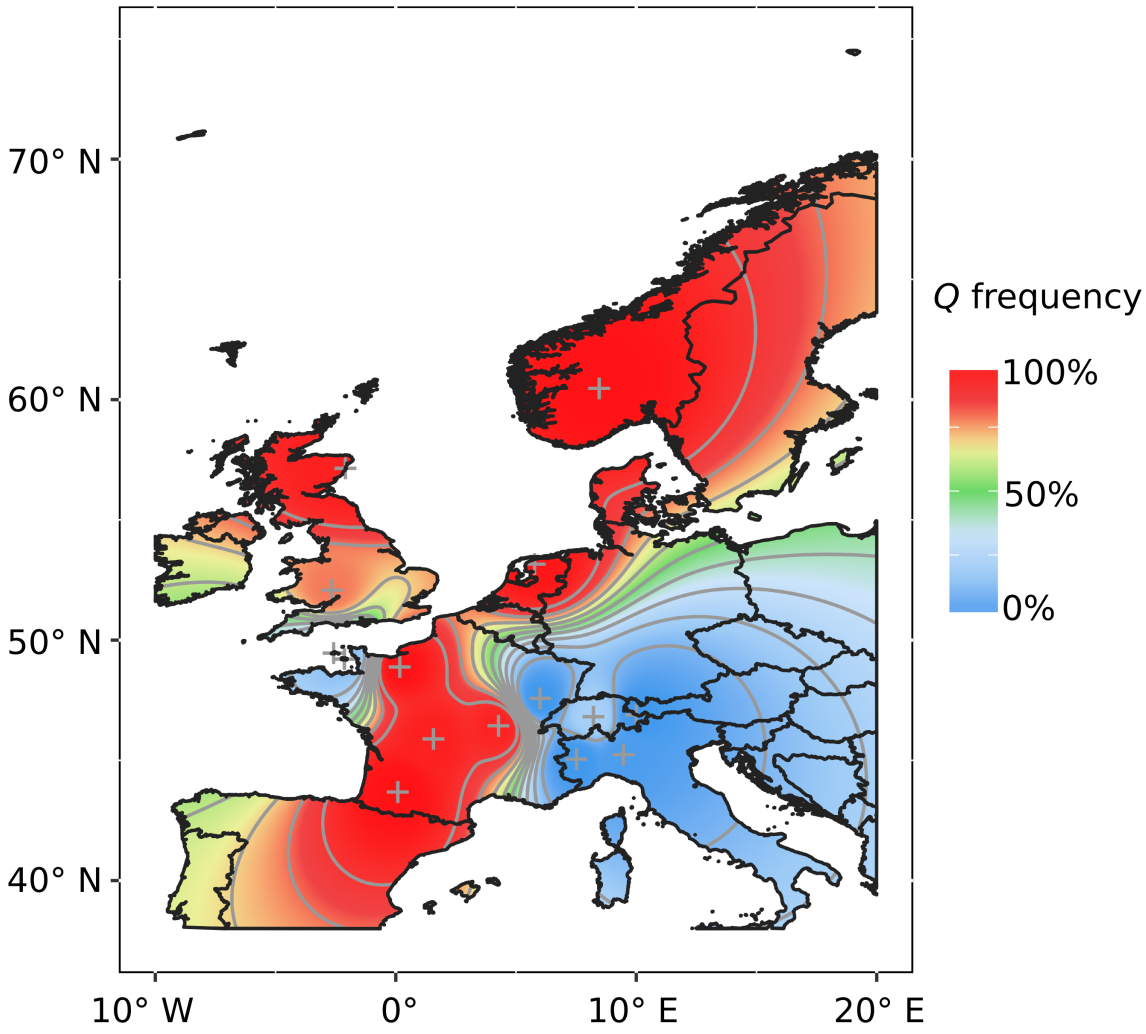

Figure 5. Atlantic Europe as the most likely centre of recent selection for *Q*. The heatmap was generated with ggplot2 v2.1.0 (WICKHAM, 2009) using inverse weighted distance exact interpolation of GGGTTCG frequency from breed origin (crosses).

3.4. Extended homozygosity and archaeological data indicate a role of *Q* in stature recovery

We used extended haplotype homozygosity (STEPHENS et al., 1998; SABETI et a., 2002) to estimate the age of the selective sweep in Northwestern European breeds (Figure 6a) and found that it dated to ~386 yBP (95% CI [305 – 475]). This corresponds to the period between the 16th and 18th centuries, which overlaps with stature recovery in European cattle (GUINTARD, 1999; AJMONE-MARSAN; GARCIA; LENSTRA, 2010). Simulations suggested that coefficients of selection between 0.12 and 0.27 would be required to rise the frequency of *Q* close to fixation in a time span between 60 (18th century) and 100 (16th century) generations (Figure 6b). These levels of selective pressure were comparable with those estimated for lactase persistence in humans (BERSAGLIERI et al., 2004), indicating strong selection at the *PLAG1* locus. One standing issue, however, is that haplotype coalescence could only provide an estimate of when the haplotype might have started to increase in frequency rapidly, which does not necessarily coincide with the age of the mutation. The identification of carriers of *Q* among animals predating the selective sweep is therefore needed in order to gain insights on the age of the mutation.

Coalescence of the GGGTTCG haplotype in our sample of Nellore bulls was estimated at ~65 yBP (95% CI [30, 100]), consistent with introgression after imports to Brazil from India in the 20th century. Likewise, coalescence in Brahman was ~121 yBP (95% CI [75, 170]), consistent with the period of formation and grading up of this breed. These results provided further evidence of *B. taurus* introgression as a source for *Q* in *B. indicus* populations. We also noticed that coalescence for haplotypes underlying the dominant mutations causing the polled phenotype had intervals matching with those found for *Q* in both Northwestern European and Nellore cattle, which suggested a parallel spread of the *Q* and the polled mutations worldwide (APPENDIX B).

The hypothesis of a Northwestern European selective sweep for *Q* could be refuted by the high frequency of GGGTTCG (53.9%) in the Japanese breed Wagyu. Nishimura et al. (2012) also reported segregation of the *Q* allele in Japanese Black

cattle, and declared an unknown origin of *Q* in Japan. However, Japanese breeds were admixed with Northwestern European cattle between 1868 and 1918 (DECKER et al., 2014). Indeed, analysis of Y chromosome haplotypes in Wagyu revealed multiple admixture events, including introgression from Northwestern cattle (APPENDIX B). Here, coalescence was younger in Wagyu than in other *B. taurus* breeds and estimated at ~141 yBP (95% CI [90, 195]), which falls in the introgression period and supports a Northwestern European origin of *Q* also in Japanese cattle.

If the intense selection for *Q* facilitated stature recovery in Northwestern Europe, a sharp increase in body size should be observed in cattle from the 16th – 18th centuries in comparison to previous centuries. To gain insights on the extent of increase in body size in this period, we analyzed size measurements of bone fragments recovered from two archaeological sites in Iceland (HAMBRECHT, 2012; HARRISON, 2014): (i) Gásir (65º 46' 58" N – 18º 9' 58" W), containing specimens from the 14th – 15th century (n = 33); and (ii) Skalholt (64º 7' 38" N – 20º 31' 35" W), containing specimens from the 17th – 18th century (n = 89). In order to make the data comparable across different bone fragments, we computed the log-ratio between measurements taken from each specimen and equivalent measurements from a reference skeleton of a 19th century cow from upstate New York kept at the American Museum of Natural History (specimen #14908, withers height of 94.9 cm). The difference in the mean log-ratio between the 17th – 18th century and 14th – 15th century data was 0.076 ± 0.019 (p = 7.04 x 10^{-5}), suggesting a ~1.2-fold increase in body size within the time frame when the frequency of *Q* arose rapidly (Figure 6c).

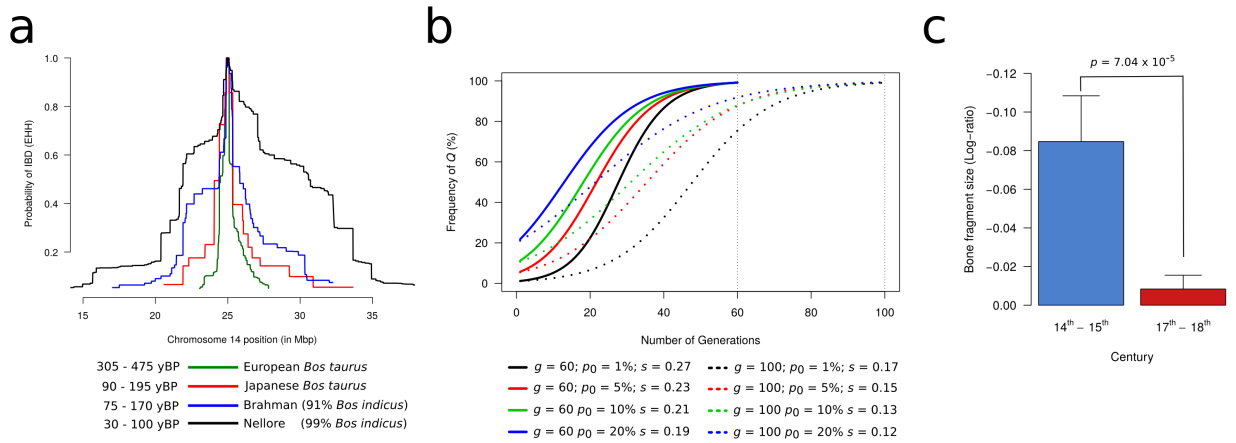

Figure 6. Time to coalescence for the *PLAG1* haplotype. (a) Extended haplotype homozygosity (EHH) analysis (rooted at rs109815800). In Northwestern European breeds, the signature dates back to the period of stature recovery in the 16th - 18th century. The haplotype is much more recent in *B. indicus* and Japanese *B. taurus* breeds, indicating introgression. (b) Simulation of increase in haplotype frequency according to different selection coefficients (s) and frequency prior to selection (p_0). (c) Mean log-ratio from bone fragments recovered from archaeological sites in Iceland dating to the 14th – 15th century and the 17th – 18th century. The later period corresponds to early selection for *Q* and presents bone fragments on average 1.2 times larger than those observed in the earlier period. Error bars represent the standard errors of the means.

3.5. Ancient DNA shows that *Q* has been segregating in *B. taurus* for at least 1,000 years

Insights into the actual age of the *Q* mutation could be gained from the analysis of ancient bovine DNA (aDNA) from Northwestern Europe. In particular, allele C at rs109231213 can be used as a tag for *Q* since this variant affects a highly constrained position (PhastCons score = 1 in 100 vertebrates) (FORTES et al., 2013a) and because it is a functional candidate given its location on the 3'-UTR of *PLAG1* (KARIM et al., 2011). We first attempted to find the C allele in short read alignments reported by Park et al. (2015) of a well-preserved *Bos primigenius* humerus bone recovered from Carsington Pasture Cave in Derbyshire, England (53º 4' 47" N – 1º 38' 7" W). This specimen was radiocarbon dated to 6,738 ± 68 yBP and therefore precedes the introduction of domesticated cattle in Britain. Inspection of rs109231213 in their alignment data revealed a clear GG genotype supported by ten single reads, indicating a *qq* genotype (Figure 7a). In spite of being homozygous wild type, this specimen did not present the ancestral haplotype at HD markers. In fact, the AGATCCT haplotype found in this sample was rare in all modern cattle breeds included in the HapMap set, except for Romagnola. Although Romagnola has been hypothesized to carry *B. indicus* introgression (MCTAVISH et al., 2013; UTSUNOMIYA et al., 2014a), AGATCCT was not found in the *B. indicus* data. This observation reinforced the occurrence of *q* in multiple haplotype backgrounds and indicated that: (i) the specimen was most likely a representative of the non-domesticated Neolithic relative of European *B. taurus* rather than a proxy for *B. primigenius* ancestral to *B. taurus* and *B. indicus*; and (ii) future studies may need to differentiate between introgression from the *B. indicus* lineage and introgression from non-domesticated *B. taurus*. The later is especially important since model-based clustering algorithms tend to cluster *B. indicus* and outgroups together when only two ancestral populations are assumed, in which case hybridization with outgroups becomes indistinguishable from *B. indicus* introgression.

Successful sampling of carriers of *Q* in assemblages of cattle remains predating the selective sweep might be challenging if *Q* was rare prior to selection. However, if our hypothesis of parallel spread of *Q* and polledness holds, DNA

extraction from polled specimens may increase the probability of obtaining a CC or CG genotype at rs109231213. Therefore, we extracted aDNA from two naturally polled crania from Iceland. The first specimen dating over 1,000 yBP (#159, Figure 7b) was recovered from a ritual gathering site in Hofstadir (65º 36' 47" N – 17º 10' 2" W) (LUCAS; MCGOVERN, 2007), and was considered rare in the sense that polledness was not a common trait for Icelandic cattle from the Viking Age (8th to 11th century). The second specimen (#2439, Figure 7c) dated over 300 yBP and was collected from the Skalholt site. Sequencing of cloned PCR products including the rs109231213 position revealed the occurrence of allele C in both specimens: the 19 clones obtained from specimen #159 consistently indicated a CC genotype (Figure 7b), whereas specimen #2439 returned eleven clones with G and four clones with C, indicating a CG genotype (Figure 7c). Therefore, it seems that both the *Q* and the *POLLED* mutations are at least 1,000 years old.

Since large Roman cattle were introduced in Northwestern Europe during the 1st century, we attempted to detect allele C in aDNA from Northern Italy to test the hypothesis that *Q* was brought from the Central/Southern to the Atlantic region of Europe by the Roman Empire. We analyzed a molar tooth (MU15) and a long bone (MU18) obtained from a late Roman site in Northeastern Italy (46º 7' 11" N – 13º 7' 24" E) dating over 1,700 yBP. Additionally, two molar teeth (US123 and US124) from different stratigraphic units and thus likely belonging to different individuals from a 13th – 14th century Medieval site in Northern Italy were also analyzed (46º 15' 54" N – 10º 26' 42" E). All 45 clones obtained from these samples exhibited GG genotypes (Figure 7d). As our sample size was fairly small, we could not discard occurrence of *Q* in Roman cattle. However, the consistency in the retrieval of GG genotypes indicated that *q* was most likely the major allele in this population.

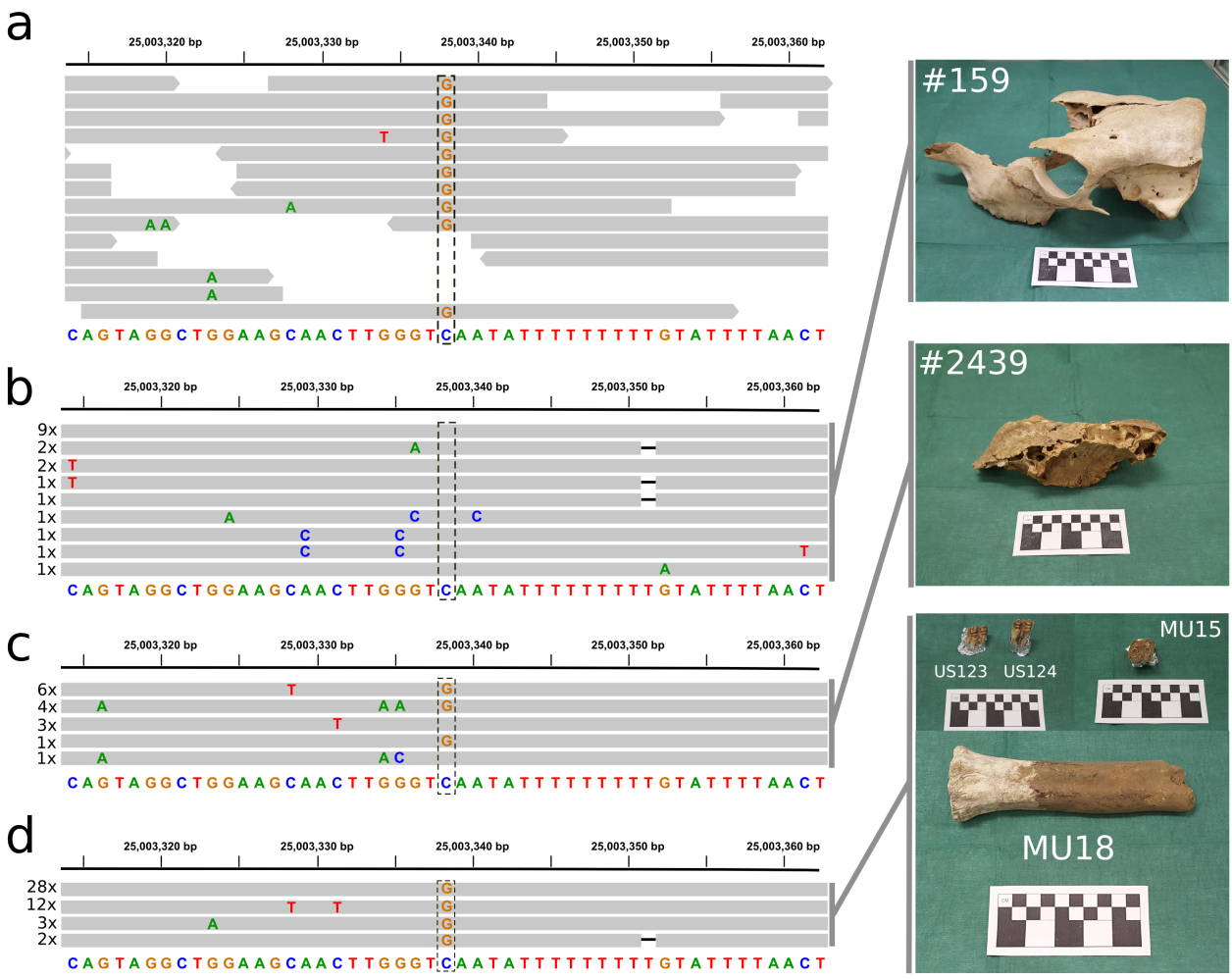

Figure 7. Insights on the age of *Q* from rs109231213 genotypes in ancient DNA. (a) Alignments of next-generation sequencing data reported by Park et al. (2015) from a >6,000 years old bovine humerus found in England. This specimen carried a *qq* genotype, as evidenced by all reads (horizontal grey bars) presenting allele G at rs109231213 (dashed rectangle). (b) A polled cranium (specimen #159) from ~1,000 yBP was recovered from a ritual gathering site in Hofstadir, Iceland. Target-sequencing of its petrous bone revealed a *QQ* genotype (i.e., all clones exhibited allele C at rs109231213). (c) A second polled cranium (specimen #2439) recovered from Skalholt and dating >300 yBP returned a heterozygous genotype. (d) Two molar teeth from a Medieval site in Northern Italy (US123 and US124) and a molar tooth (MU15) and a long bone (MU18) belonging to the late Roman age (>1,700 yBP) all presented *qq* genotypes.

4. Discussion

In the present study we provided evidence that the selective sweep encompassing *PLAG1* dates back to the period of stature recovery in Northwestern European cattle (16th – 18th century) (GUINTARD, 1999). Also, geographical distribution of *Q* followed a pattern opposite to that of stature loss (LASOTA-MOSKALEWSKA; KOBRYN, 1990), reinforcing a role for this mutation in stature recovery. The centre of the selection event for *Q* seems to lie in Atlantic Europe. The actual *Q* mutation is at least one millennium old, as suggested by sequencing of an Icelandic *B. taurus* cranium from ~1,000 yBP.

From its putative centre of origin, our analysis suggested that the *Q* allele spread towards Central, Southern and Eastern Europe, but the intensive selection for pure lines in the past few centuries may have constrained its prevalence in these areas. The *Q* allele was most likely introduced into the New World by colonizers via imports from Northwestern Europe (AJMONE-MARSAN; GARCIA; LENSTRA, 2010). Cattle trades between Northwestern Europe and Japan in the late 19th and early 20th century (DECKER et al., 2014) probably favored an introgression of *Q* to Japanese breeds. More recently, *Q* was introgressed into *B. indicus* breeds. For instance, *Q* is speculated to have favored the 'grading up' of Brahman cattle, which consisted in introgressing major *B. taurus* alleles into the breed by crossbreeding and back-crossing (FORTES et al., 2013a). Another example provided by our study is Nellore, which gained the *Q* allele after imports to Brazil from crossbreeding with cattle descended from European breeds. This introgression most likely occurred during the population expansion in the late 19th and early 20th century.

In spite of our exhaustive analysis, the identity of *Q* remained unknown. Karim et al. (2011) reported that none of the eight *B. taurus* positional candidates resided in coding regions or were associated with genotype-specific gene products. However, the authors found haplotype-dependent differences in fetal expression of seven genes, notably *PLAG1*, suggesting a regulon affected by a *cis*-acting causal nucleotide. They further used allelic imbalance and luciferase reporter assays to suggest causality of one of the following three regulatory variants (Figure 8): (i) rs109231213 (CHR14:25003338), a C to G substitution; (ii) ss319607405

(CHR14:25052396-25052398, also rs209821678), a tandem repeat of eleven or nine CCG copies; and (iii) ss319607406 (CHR14:25052440, also rs210030313), a G to A substitution. All of the three variants changed highly conserved nucleotides across mammals: rs109231213 (PhastCons score = 1) is located at the 3'-UTR of *PLAG1*, whereas ss319607405 (PhastCons score = 0.999) and ss319607406 (PhastCons score = 0.992) are in the bi-directional promoter of coiled-coil-helix-coiled-coil-helix domain containing 7 gene (*CHCHD7*) and *PLAG1*. Still, the individual contributions of these and other variants to the haplotype effect remained unclear.

Interestingly, rs109231213 was the only previously reported functional candidate included in our haplotype region. If not the causal variant itself, our analysis at least suggested that rs109231213 is a reliable proxy. In fact, Fortes et al. (2013a) have successfully used alleles C and G at rs109231213 as tags for the *Q* and *q* alleles in Brahman, respectively. This is convenient because the other two functional candidates are difficult to type, since they are located in a repetitive GC-rich region. For instance, Nishimura et al. (2012) reported difficulties in amplifying the *PLAG1-CHCHD7* promoter for sequencing. In the recent selection signatures study by Boitard et al. (2016), variants ss319607405 and ss319607406 were not included due to low quality scores and alignment issues. Altogether, these results indicate that in the absence of knowledge about the causal variant, the GGGTTCG haplotype from the HD array or the C allele at SNP rs109231213 could be used to analyze *Q* segregation, trait-association, gene flow, selection and drift in worldwide cattle.

The quest for the identity of *Q* continues. Variants ss319607405 and ss319607406 are appealing candidates since they were shown to change transcriptional activity of the bi-directional promoter of *PLAG1* and *CHCHD7* (KARIM et al., 2011). However, rs109231213 is also a strong candidate since it affects a highly conserved nucleotide at the 3'-UTR of *PLAG1*. Variation at the 3′-UTR of a transcript may affect its interaction with regulatory molecules, such as MicroRNAs (miRNA) (ARNOLD et al., 2012), and therefore impact levels of translation. Expression levels of *PLAG1* were previously shown to be regulated by miRNAs binding to this 3'-UTR region (PALLASCH et al., 2009; PATZ; PALLASCH; WENDTNER, 2010), which also supports candidacy of rs109231213. Indeed, the G to C substitution at rs109231213 is predicted to cause the loss of a highly conserved

miRNA binding-site. Another hypothesis involves shared causality of the entire haplotype, meaning *Q* = C(CCG)$_{11}$G and *q* = G(CCG)$_{9}$A, with potential epistatic effects. Separating these effects and solving the causality at this locus will demand the analysis of recombinants at these three loci, which are difficult to find naturally. A recombinant construct between ss319607405 and ss319607406 showed that the (CCG)$_{11}$G haplotype is necessary to change transcription activity (KARIM et al., 2011), which suggests interaction between the two loci. In order to disentangle interactions between rs109231213 and the other two variants, gene-edited cell lines of G(CCG)$_{11}$G or C(CCG)$_{9}$A recombinants are required and should be the focus of future studies. Moreover, a complex mutation comprising cooperation of multiple regulatory variants at the haplotype region is not yet to be discarded, since other six genes besides *PLAG1* have their expression levels affected to some degree by *Q*. Yet, *PLAG1* seems to be the leading gene behind most of the pleiotropic effects observed at this chromosomal domain, as suggested by a recent expression QTL study (FINK et al., 2017). Finally, previously unavailable functional annotations such as expression QTLs, promoters, enhancers, insulators, transcription factor binding sites, CTCF loops and 3D chromatin structure, among others, will be soon available from the upcoming *B. indicus* assembly, the new long-read *B. taurus* assembly and the Functional Annotation of Animal Genomes (FAANG) projects. As our understanding of regulatory elements in the bovine genome progresses, other candidate variants may also emerge and reveal additional trait-specific effects.

In conclusion, we were able to demonstrate that the pleiotropic QTL and signature of selection spanning *PLAG1* are most likely explained by a single mutation (or haplotype) across modern worldwide cattle breeds. We also found that this mutation has been selected during the period of stature recovery in Northwestern Europe and was later spread around the globe via importation of cattle from Atlantic Europe. Altogether, this study presents one of the first examples of a selective sweep in livestock that was driven by strong (supposedly artificial) selection on a complex trait.

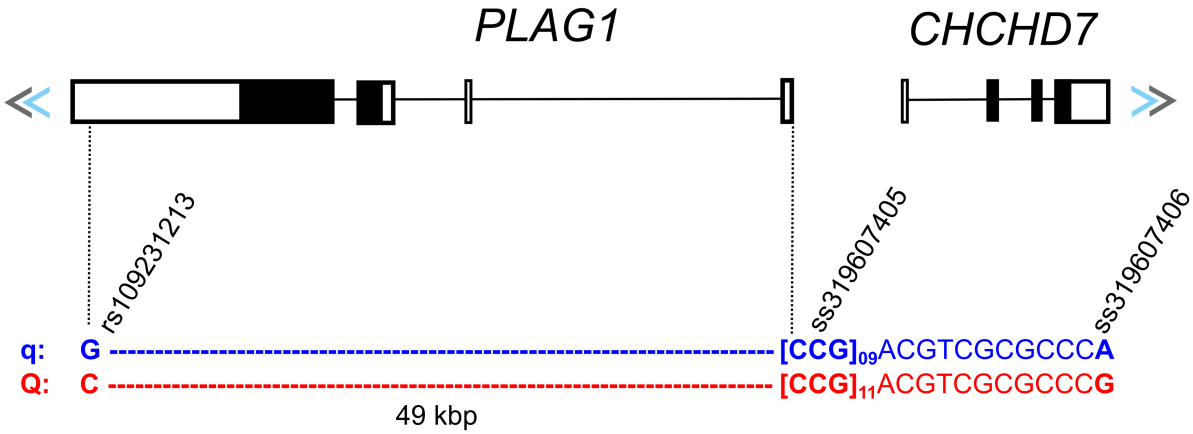

Figure 8. Functional candidate variants underlying the *PLAG1* chromosomal domain. Variant rs109231213 is predicted to change a conserved miRNA binding site, whereas variants ss319607405 and ss319607406 modify the transcription activity of the bi-directional promoter of *PLAG1* and *CHCHD7.*

5. Acknowledgments

This research was funded by Sao Paulo Research Foundation (FAPESP, process 2010/52030-2, 2014/01095-8 and 2016/07531-0), National Council for Scientific and Technological Development (CNPq, process 560922/2010-8 and 483590/2010-0) and Coordination for the Improvement of Higher Education Personnel (CAPES).

6. Author contributions

J.F.G., Y.T.U., Ta.S.S. and J.S. conceived and designed the study. J.F.G., Ta.S.S., C.P.V.T., A.S.C., S.S., D.B. and G.L. coordinated sequencing of *B. indicus* bulls. J.F.G., Ta.S.S., C.P.V.T. and A.S.C. coordinated genotyping of *B. indicus* bulls. R.C. and H.H.R.N. provided birth weight data. R.C.P., Th.S.S., L.B.Z., R.S.C. and M.M.T.C. scored horned/polled phenotypes. L.C., E.E. and P.A.M. extracted ancient DNA and conducted targeted sequencing for the ancient samples. G.H. contributed Icelandic archaeological samples and data. D.C., M.B., Mi.M. and M.S. contributed Italian archaeological samples and data. Y.T.U., Ma.M. and R.B.P.T. performed

haplotype analyses. Y.T.U., A.T.H.U., M.S.C. and T.S.A. performed association analyses. Y.T.U., Ma.M., E.K. and R.B.P.T. performed sequence data analyses. Y.T.U. wrote the manuscript. All authors revised and agreed with the contents of the manuscript.

7. Competing financial interests

H.H.R.N. is employed by Gensys Consultores Associados and Ta.S.S. is employed by Recombinetics, Inc. All other authors declare no potential conflicts of interest. 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. Data availability

The data that support the findings of this study were obtained under license and so are not publicly available. Data are however available for academic use from the authors upon reasonable request.

9. Methods

9.1. Phasing and haplotyping

Nellore data were filtered in PLINK v1.90 (PURCELL et al., 2007; CHANG et al., 2015) for a minium call rate of 95% and minor allele frequency of at least 5%. A total of 447,617 markers (out of 786,799) were retained for analysis. Genotypes were phased with the Segmented HAPlotype Estimation & Imputation Tool v2.r837 (SHAPEIT2) (O'CONNELL et al., 2014). The following parameter values were used: effective population size (Ne_{LD}) of 113, burn in of ten iterations, prune of ten iterations, 50 main iterations, 200 states, and windows of 500 kbp. Effective population size was estimated from the data with SNeP v1.1 (BARBATO et al., 2015). Phasing with default parameters and a previously published value for Ne_{LD} of 362 (ZAVAREZ et al.,

2015) based on a larger sample of Nellore animals yielded very similar results (data not shown). We used GHap v1.2.2 (UTSUNOMIYA et al., 2016b) to determine haplotype alleles within chromosomal segments and score haplotype genotypes for each animal. The segment size was chosen based on the extent of LD in the Nellore genome and the intermarker spacing of the HD panel. Given that markers were placed on average every ~5 kbp on CHR14, and that LD extends up to 30 kbp in the Nellore genome (ESPIGOLAN et al., 2013), we determined haplotypes within overlapping segments of six consecutive markers. A total of 15,127 sliding windows of six consecutive markers were screened throughout autosome 14 for short haplotype calling. The same phasing and short haplotype calling procedures were later applied to CHR1 and HapMap genotypes, with Ne_{LD} reset to the default value of 10,000 in the later case.

9.2. Phenotypes

Birth weight estimated breeding values were obtained from a single-trait animal model fitted to records from 846,782 calves born between 1985 and 2012 in 315 grazing-based Brazilian herds. The model included the fixed effects of contemporary group (defined as animals from the same herd, born in the same year and season, and belonging to the same birth management group) and age of dam at calving, as well as random maternal effects (maternal additive genetic effect and maternal permanent environmental effect) (UTSUNOMIYA et al., 2013). The heritability was estimated at 0.37. Prior to the association analysis, estimated breeding values were deregressed following Garrick, Taylor and Fernando (2009). The minimum, mean, standard deviation and maximum of the accuracy of dEBV (based on prediction error variance) was 0.58, 0.95, 0.30 and 0.98, respectively. Presence or absence of horns was scored by five independent evaluators from pictures of the genotyped Nellore bulls. Majority voting was used to assign phenotypes to animals, and at least three voters were required to declare an animal to be horned or polled. Furthermore, pictures from animals suspected to be surgically de-horned were excluded. The resulting data included 328 horned and 51 polled bulls.

9.3. Haplotype regression analysis

The parameterization presented here builds on previously reported models (FALCONER, 1960; DA, 2015). For a given window, let haplotype alleles 1, 2, …, *K* be sorted by frequencies p_1, p_2, …, p_K. We start by defining a_k as the average effect of substituting the minor haplotype allele 1 by k = 2, 3, …, *K*. From these definitions, the partial breeding value associated with haplotype *k* is:

$$u_k = z_k\alpha_k$$

where z_k is a scalar taking values:

$0 - 2p_k$, for 0 copies of haplotype *k*
$1 - 2p_k$, for 1 copy of haplotype *k*
$2 - 2p_k$, for 2 copies of haplotype *k*

phenotypes were regressed onto haplotype-specific breeding values using the model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

where **y** is the vector of phenotypes, **b** = [m α_k]', m is an intercept, **X** = [**1** $\mathbf{z}_k$], **1** is a vector of ones, $\mathbf{z}_k$ is a vector relating phenotypes to substitution effect α_k, and **e** is a vector of residuals distributed as $N(0,\mathbf{W}\sigma^2_e)$, where **W** = Diag(**w**), **w** is a vector of weights, and σ^2_e is the residual variance. In the case of birth weight dEBVs, the vector of weights was defined as $\mathbf{w} = \lambda^{-1}(\mathbf{d}+c)$ (GARRICK; TAYLOR; FERNANDO, 2009), where $\lambda = (1 - h^2)/h^2$, h^2 is the heritability of birth weight, $\mathbf{d} = (1 - \mathbf{r}^2)/\mathbf{r}^2$, $\mathbf{r}^2$ is the vector of dEBV reliabilities, and *c* is the assumed proportion of genetic variance for which haplotypes cannot account. We have previously estimated (UTSUNOMIYA et al., 2013) a 95% confidence interval of 2.12-8.09% for the genetic variance due to the targeted QTL. Therefore, we adopted a value of *c* = 0.9191 (i.e., the candidate haplotype can only explain at most 8.09% of the genetic variance). For the

polledness analysis, horned and polled animals had phenotypes coded as -1 and 1, respectively, and **W** was replaced by an identity matrix. Regression was carried out by solving the generalized least squares equations $\mathbf{b} = (\mathbf{X'W^{-1}X})^{-1}\mathbf{X'W^{-1}y}$. Association mapping was then based on the two-tailed t-test α_k / SE(α_k), and the most significant haplotype was selected as a tag for the underlying causal mutation. Similar results were found when association tests were performed using haplotype windows of 1 (i.e., single SNP), 5, 10 and 20 markers with correction for polygenic effects (APPENDIX B).

9.4. Analysis of ancestry and genetic structure

The unsupervised model-based clustering algorithm in Admixture 1.3 (ALEXANDER; NOVEMBRE; LANGE, 2009) was used to estimate *B. taurus* ancestry on CHR14 in Nellore cattle. The HapMap data was merged with the Nellore data and the algorithm was run assuming two ancestral populations (*B. taurus* and *B. indicus*). Admixed and outgroup populations were excluded from this analysis. Principal components were computed with PLINK v1.90 (PURCELL et al., 2007; CHANG et al., 2015). Statistical tests for differences in *B. taurus* ancestry on CHR14 and mean birth weight dEBV between breeding subgroups of Nellore cattle were based on a Mann-Whitney U test and a t-test, respectively. Differences in frequency of *Q* and polledness were assessed with Fisher's exact test. Local ancestry was estimated using the same reference data with elai v1.00 (GUAN, 2014). Ten independent runs with different random seeds were performed in parallel, from which eight were based on the Expectation-Maximization (EM) algorithm and two were obtained using a fast linear approximation. The EM algorithm was run with 30 steps. All analyses assumed ten generations since admixture and numbers of upper and lower clusters of two and ten, respectively. The final output was built from the average across the ten replicates.

9.5. Extended haplotype homozygosity

For each marker pertaining to the haplotype of interest, and considering only

chromosomes carrying that haplotype, the decay in the probability of identity-by-descent was modeled using extended haplotype homozygosity (EHH), following Sabeti et al. (2002). Briefly, EHH between a core marker and another marker upstream or downstream was computed as:

$$EHH = \frac{\sum_{h=1}^{H} \binom{n_h}{2}}{\binom{N}{2}}$$

where (:) denotes the binomial coefficient, n_h is the count of haplotype *h* carrying the allele of interest and *N* is the total number of chromosomes carrying that allele. Calculation of EHH was repeated for all markers upstream and downstream of a core marker until EHH decayed to 0.05, corresponding to a 5% probability of obtaining a pair of haplotypes identical by descent when two chromosomes were sampled at random. All EHH calculations were performed using the *rehh* v2.0.0 R package (GAUTIER; VITALIS, 2012). The expected number of generations until coalescence was obtained as (STEPHENS et al., 1998):

$$E[g] = 1 - \ln(p)/r$$

where *p* is the probability of identity-by-descent and *r* is the segment size in Morgans. The latter was computed from the entire length span of the EHH decay assuming 1 cM ~ 1.23 Mbp (MA et al., 2014). The use of a genetic map to account for local recombination rates yielded very similar results (APPENDIX B). Time to haplotype coalescence was computed based on a cattle generation interval of five years. Confidence intervals were derived from 1,000,000 Monte Carlo simulations assuming that the number of generations followed a Poisson distribution with parameter $\lambda = E[g]$. These simulations were verified to be equivalent to assuming estimate error $e \sim N(0, \lambda)$.

9.6. Coefficient of selection

The magnitude of selection required to fix the *Q* mutation in the course of 60 or 100 generations was computed considering the equation relating average fitness in a population (*w*) to allele frequencies under constant selection (STEPHENS et al., 1998):

$$w = p^2 w_{QQ} + 2p(1 - p)w_{Qq} + (1 - p)^2 w_{qq}$$

where p is the frequency of *Q*, $w_{QQ} = 1$, $w_{Qq} = 1 - hs$ and $w_{qq} = 1 - s$ are the relative fitness of genotypes *QQ*, *Qq* and *qq*, respectively, h is a scalar taking values 0, 0.5 or 1 if the effect of *Q* was assumed dominant, additive or recessive, respectively, and s is the coefficient of selection. Based on the additive distribution of phenotypes according to genotypes, we adopted $h = 0.5$. Change in p from one generation to the next was then computed as:

$$p_{t+1} = p_t[p_t w_{QQ} + (1 - p_t)w_{Qq}]/w$$

We simulated a range of scenarios with different frequencies for *Q* at $t = 0$, starting at 1% and going up to 20%. For each scenario, values of s ranging from 0.01 to 0.50 were tested in order to identify a coefficient of selection sufficient to increase the frequency of *Q* up to 99% in 60 or 100 iterations.

9.7. Analysis of bone measurements

The archaeological data used for size analysis comprised fragments of humerus (n = 38), tibia (n = 27), metatarsal (n = 24), metacarpal (n = 11), calcaneus (n = 2), femur (n = 15) and radius-ulna (n = 5) bones. According to bone type and integrity, measurements included breadth of the distal end (Bd, n = 14), breadth of the proximal end (Bp, n = 13), breadth of trochlea (BT, n = 13), depth across the processus anconaeus (DPA, n = 10), greatest breadth (GB, n = 15), greatest length (GL, n = 18), smallest breadth of diaphysis (SD, n = 16) and smallest depth of

olecranon (SDO, n = 12). Equivalent measurements were taken from a 19th century reference cow skeleton (specimen #14908 from the American Museum of Natural History) and the ratio between specimen and reference was calculated. Ratios were then transformed to a log10 scale. A t-test was used to compare the log-ratio means between specimens from the 17th – 18th century and the 14th – 15th century.

9.8. Analysis of ancient DNA

A total of seven samples were used for aDNA analysis: two petrous bones from a 10th century bovine cranium (specimen #159) from the site of Hofstadir and one petrous bone from a 15th – 17th century bovine cranium (specimen #2439) from the site of Skalholt; one molar tooth (MU15) and one long bone (MU18) from the 3rd – 4th century from the late Roman site of Muris di Moruzzo (UD) in Northeastern Italy (46º 7' 11" N – 13º 7' 24" E); and two molar teeth (US123 and US124) from different stratigraphic units and thus likely belonging to different individuals from the 13th – 14th century Medieval site of Tor dei Pagà (Vione, BS) in Northern Italy (46º 15' 54" N – 10º 26' 42" E). Extraction, PCR amplification and cloning of aDNA were performed in a dedicated lab facility of the BioDNA Research Centre (Piacenza, Italy). To exclude contamination from exogenous sources, stringent criteria for aDNA analysis were followed: extraction and PCR set up were carried out in physically separated clean rooms, irradiated with UV light (254 nm wavelength) each night for 2h and additionally after every work session, and equipped with a positive air pressure system to prevent external contaminants from entering. All post-PCR analyses were performed in a separate modern-DNA lab. Disposable Tyvek® coveralls and double gloves were worn during the experiments and were changed frequently. All benches and rooms were routinely treated with bleach and UV-light. Extraction and amplification blanks were used as negative controls in each experiment. In the “extraction” clean room, after a 2h UV light preliminary irradiation of the samples, the external layer of the surface was removed with disposable burs on a Marathon Multi 600 dental device. The samples were further UV-irradiated for 45 minutes and ground into powder with the same tool. Starting from about 250 mg of teeth/bone powder per sample, DNA was extracted according to the method of Rohland and Hofreiter

(2007). Amplification reactions were assembled in the “PCR set up” clean room under a Biosan UVC/T-AR cabinet. PCR amplification of a 104 bp fragment flanking the target polymorphism (rs109231213) was obtained with the primer pair Bt_PLAG1_F 5’ CTCAAAACACACTGTCTTCCCA 3’ and Bt_PLAG1_R 5’ GATCTCCTCCAATGTCGCCT 3’. Two ml of extracted DNA were added to 50 ml PCR reaction mix including 2U of AmpliTaq Gold® DNA Polymerase (Applied Biosystems), 2 mM $MgCl_2$, 160 µM of each dNTP and 1 µM of each primer. The amplification thermal profile was as follows: 95°C for 10 min (polymerase activation), 45 cycles of denaturation, 95°C for 20 sec, annealing, 50°C for 20 sec and extension, 72°C for 30 sec, and final extension at 72°C for 10 min. PCR products were checked on 1.5% agarose gel. Bands of expected size were excised from agarose gel, purified with Sepharose® Cl-6B resin (GE Healthcare) on Micro Biospin Columns 732-6204 (BIORAD) and cloned with the TOPO® TA Cloning® Kit (Invitrogen™) following the manufacturer’s instructions. A total of 10 to 15 clones per sample were amplified from white recombinant colonies with universal M13 primers. After purification with Sepharose® resin, PCR products were sequenced with M13 forward primer in outsourcing at Macrogen company (http://www.macrogen.com). The resulting chromatograms were visually inspected with SeqTrace v0.9.0 (STUCKY, 2012) and aligned to the *PLAG1* sequence of the cattle reference genome with BioEdit v7.2.5 (HALL, 1999).

9.9. Analysis of next-generation sequencing data

Paired-end libraries of 24 Nellore bulls were sequenced in the Illumina® HiSeq2000 platform, following the manufacturer's protocol (available at: http://support.illumina.com/sequencing/sequencing_instruments/hiseq_2000.html). Reads were aligned against the UMD v3.1 *B. taurus* assembly (ZIMIN et al., 2009) using the Burrows-Wheeler Alignment (BWA) algorithm v0.7.10-r789 (LI; DURBIN, 2009). After alignment, optical and PCR duplicates were marked with PicardTools v1.119 (available at: http://picard.sourceforge.net). The total sequencing yield was 2 x 3,205,981,178 paired-end reads of 100 bases (~641 billion bases). Considering only properly paired reads, 92.2% were successfully mapped, resulting in an average fold

coverage per sample of 9.25±1.48x, with a minimum of 5.87x and a maximum of 13.75x. The overall percentage of optical/PCR duplicates was 7.0%. Variants on CHR14 were extracted from aligned reads using the mpileup algorithm from SAMtools v1.3.1 and BCFtools v1.3.1 (LI et al., 2009), following guidelines for cattle data reported by Baes et al. (2014). Variant effects were predicted and annotated with Ensembl Variant Effect Predictor (VEP) (MCLAREN et al., 2016). Sequence alignments were visually inspected to confirm variant positions with both Integrative Genomics Viewer (IGV) v2.3 tool (ROBINSON et al., 2011; THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013) and the SAMtools v1.3.1 tview application (LI et al., 2009).

10. References

ACHILLI, A.; OLIVIERI, A.; PELLECCHIA, M.; UBOLDI, C.; COLLI, L.; AL-ZAHERY, N.; ACCETTURO, M.; PALA, M.; KASHANI, B. H.; PEREGO, U. A.; BATTAGLIA, V.; FORNARINO, S.; KALAMATI, J.; HOUSHMAND, M.; NEGRINI, R.; SEMINO, O.; RICHARDS, M.; MACAULAY, V.; FERRETTI, L.; BANDELT, H. J.; AJMONE-MARSAN, P.; TORRONI, A. Mitochondrial genomes of extinct aurochs survive in domestic cattle. **Current Biology**, v. 18, p. 157-158, 2008.

AJMONE-MARSAN, P.; GARCIA, J. F.; LENSTRA, J. A. On the origin of cattle: how aurochs became cattle and colonized the world. **Evolutionary Anthropology: Issues, News, and Reviews**, v. 19, p. 148-157, 2010.

ALEXANDER, D. H.; NOVEMBRE, J.; LANGE, K. Fast model-based estimation of ancestry in unrelated individuals. **Genome Research**, v. 19, p. 1655-1664, 2009.

ARNOLD, M.; ELLWANGER, D. C.; HARTSPERGER, M. L.; PFEUFER, A.; STÜMPFLEN, V. Cis-acting polymorphisms affect complex traits through modifications of MicroRNA regulation pathways. **PLOS ONE**, v. 7, e36694, 2012.

BAES, C. F.; DOLEZAL, M. A.; KOLTES, J. E.; BAPST, B.; FRITZ-WATERS, E.;

JANSEN, S.; FLURY, C.; SIGNER-HASLER, H.; STRICKER, C.; FERNANDO, R.; FRIES, R.; MOLL, J.; GARRICK, D. J.; REECY, J. M.; GREDLER, B. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. **BMC Genomics**, v. 15, 948, 2014.

BARBATO, M.; OROZCO-TERWENGEL, P.; TAPIO, M.; BRUFORD, M. W. SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. **Frontiers in Genetics**, v. 6, 109, 2015.

BERSAGLIERI, T.; SABETI, P. C.; PATTERSON, N.; VANDERPLOEG, T.; SCHAFFNER, S. F.; DRAKE, J. A.; RHODES, M.; REICH, D. E.; HIRSCHHORN, J. N. Genetic signatures of strong recent positive selection at the lactase gene. **American Journal of Human Genetics**, v. 74, p. 1111-1120, 2004.

BOITARD, S.; BOUSSAHA, M.; CAPITAN, A.; ROCHA, D.; SERVIN, B. Uncovering adaptation from sequence data: Lessons from genome resequencing of four cattle breeds. **Genetics**, v. 203, p. 433-450, 2016.

BRUFORD, M. W.; BRADLEY, D. G.; LUIKART, G. DNA markers reveal the complexity of livestock domestication. **Nature Reviews Genetics**, v. 4, p. 900-910, 2003.

CHANG, C. C.; CHOW, C. C.; TELLIER, L. C.; VATTIKUTI, S.; PURCELL, S. M.; LEE, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. **GigaScience**, v. 4, 7, 2015.

DA, Y. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. **BMC Genetics**, v. 16, 144, 2015.

DECKER, J. E.; MCKAY, S. D.; ROLF, M. M.; KIM, J. W.; MOLINA ALCALÁ, A.; SONSTEGARD, T. S.; HANOTTE, O.; GÖTHERSTRÖM, A.; SEABURY, C. M.;

PRAHARANI, L.; BABAR, M. E.; CORREIA DE ALMEIDA REGITANO, L.; YILDIZ, M. A.; HEATON, M. P.; LIU, W. S.; LEI, C. Z.; REECY, J. M.; SAIF-UR-REHMAN, M.; SCHNABEL, R. D.; TAYLOR, J. F. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. **PLOS Genetics**, v. 10, e1004254, 2014.

ESPIGOLAN, R.; BALDI, F.; BOLIGON, A. A; SOUZA, F. R.; GORDO, D. G.; TONUSSI, R. L.; CARDOSO, D. F.; OLIVEIRA, H. N.; TONHATI, H.; SARGOLZAEI, M.; SCHENKEL, F. S.; CARVALHEIRO, R.; FERRO, J. A; ALBUQUERQUE, L. G. Study of whole genome linkage disequilibrium in Nellore cattle. **BMC Genomics**, v. 14, 305, 2013.

FALCONER, D. S. Values and means. In: _____. **Introduction to quantitative genetics**. New York, Ronald Press Co, 1960. cap. 7, p. 112-125.

FINK, T. A.; TIPLADY, K.; LOPDELL, T.; JOHNSON, T.; SNELL, R. G.; SPELMAN, R. J.; DAVIS, S. R.; LITTLEJOHN, M. D. Functional confirmation of *PLAG1* as the causative gene underlying major pleiotropic effects on liveweight and milk characteristics. **Scientific Reports**, v. 7, 44793, 2017.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS (FAO). **DAD-IS: Domestic Animal Diversity Information System**. Disponível em: <http://dad.fao.org/>. Acesso em: 4 set. 2017.

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**, v. 44, p. 636-647, 2013a.

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, 2012.

FORTES, M. R. S.; REVERTER, A.; KELLY, M.; MCCULLOCH, R.; LEHNERT, S. A. Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. **Andrology**, v. 1, p. 644-650, 2013b.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution (GSE)**, v. 41, 55, 2009.

GAUTIER, M.; VITALIS, R. Rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. **Bioinformatics**, v. 28, p. 1176-1177, 2012.

GUAN, Y. Detecting structure of haplotypes and local ancestry. **Genetics**, v. 196, p. 625-642, 2014.

GUINTARD, C. On the size of the ure-ox or aurochs (*Bos primigenius* Bojanus, 1827). In: WENIGER, G.-C. (Ed.). **Archäologie und Biologie des Auerochsen**. [S.l.]: Nean-derthal Museum, Mettmann, 1999. cap. 1, p. 7-21.

HALL, T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. **Nucleic Acids Symposium Series**, v. 41, p.95-98, 1999.

HAMBRECHT, G. Zooarchaeology and Modernity in Iceland. **International Journal of Historical Archaeology**, v. 16, p. 472-487, 2012.

HARRISON, R. Connecting the Land to the Sea at Gásir. In: HARRISSON, R.; MAHER, R. A. (Ed.). **Human Ecodynamics in the North Atlantic: A Collaborative Model of Humans and Nature through Space and Time**. 1. ed. London: Lexington

Books, 2014. cap. 7, p. 117-137.

HARTATI, H.; UTSUNOMIYA, Y. T.; SONSTEGARD, T. S.; GARCIA, J. F.; JAKARIA, J.; MULADNO, M. Evidence of *Bos javanicus* x *Bos indicus* hybridization and major QTLs for birth weight in Indonesian Peranakan Ongole cattle. **BMC Genetics**, v. 16, 75, 2015.

HARTIGAN, J. A.; WONG, M. A. A K-Means Clustering Algorithm. **Applied Statistics**, v. 28, p. 100-108, 1979.

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 and Differentiation**, v. 46, p. 459-470, 2004.

JUMA, A. R.; DAMDIMOPOULOU, P. E.; GROMMEN, S. V. H.; VAN DE VEN, W. J. M.; DE GROEF, B. Emerging role of *PLAG1* as a regulator of growth and reproduction. **Journal of Endocrinology**, v. 228, p. R45-R56, 2016.

KARIM, L.; TAKEDA, H.; LIN, L.; DRUET, T.; ARIAS, J. A. C.; 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, p. 405-413, 2011.

LASOTA-MOSKALEWSKA, A.; KOBRYN, H. The size of aurochs skeletons from Europe and Asia in the period from the Neolithic to the Middle Ages. **Acta Theriologica**, v. 35, p. 89-109, 1990.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, p. 1754-1760, 2009.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, p. 2078-2079, 2009.

LITTLEJOHN, M.; GRALA, T.; SANDERS, K.; WALKER, C.; WAGHORN, G.; MACDONALD, K.; COPPIETERS, W.; GEORGES, M.; SPELMAN, R.; HILLERTON, E.; DAVIS, 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, p. 591-594, 2012.

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 Sciences of the United States of America (PNAS USA)**, v. 91, p. 2757-2761, 1994.

LUCAS, G.; MCGOVERN, T. Bloody slaughter: ritual decapitation and display at the Viking settlement of Hofstadir, Iceland. **European Journal of Archaeology**, v. 10, p. 7–30, 2007.

MA, L.; O'CONNELL, J. R.; VANRADEN, P. M.; SHEN, B.; PADHI, A.; SUN, C.; BICKHART, D. M.; COLE, J. B.; NULL, D. J.; LIU, G. E.; DA, Y.; WIGGANS, G. R. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. **PLOS Genetics**, v. 11, e1005387, 2015.

MCLAREN, W.; GIL, L.; HUNT, S. E.; RIAT, H. S.; RITCHIE, G. R. S.; THORMANN, A.; FLICEK, P.; CUNNINGHAM, F. The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, 122, 2016.

MCTAVISH, E. J.; DECKER, J. E.; SCHNABEL, R. D.; TAYLOR, J. F.; HILLIS, D. M. New World cattle show ancestry from multiple independent domestication events. **Proceedings of the National Academy of Sciences of the United States of America (PNAS USA)**, v. 110, p. E1398-1406, 2013.

MURRAY, C.; HUERTA-SANCHEZ, E.; CASEY, F.; BRADLEY, D. G. Cattle demographic history modeled from autosomal sequence variation. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 365, p. 2531-2539, 2010.

NEVES, H. H. R.; CARVALHEIRO, R.; PÉREZ-O'BRIEN, A. M.; UTSUNOMIYA, Y. T.; CARMO, A. S.; SCHENKEL, F. S.; SÖLKNER, J.; MCEWAN, J. C.; VAN TASSELL, C. P.; COLE, J. B.; SILVA, M. V. G. B.; QUEIROZ, S. A.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genomic predictions in *Bos indicus* (Nellore) cattle. **Genetics Selection Evolution (GSE)**, v. 46, 17, 2014.

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**, v. 13, 40, 2012.

O'CONNELL, J.; GURDASANI, D.; DELANEAU, O.; PIRASTU, N.; ULIVI, S.; COCCA, M.; TRAGLIA, M.; HUANG, J.; HUFFMAN, J. E.; RUDAN, I.; MCQUILLAN, R.; FRASER, R. M.; CAMPBELL, H.; POLASEK, O.; ASIKI, G.; EKORU, K.; HAYWARD, C.; WRIGHT, A. F.; VITART, V.; NAVARRO, P.; ZAGURY, J. F.; WILSON, J. F.; TONIOLO, D.; GASPARINI, P.; SORANZO, N.; SANDHU, M. S.; MARCHINI, J. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. **PLOS Genetics**, v. 10, e1004234, 2014.

PALLASCH, C. P.; PATZ, M.; YOON, J. P.; HAGIST, S.; EGGLE, D.; CLAUS, R.; DEBEY-PASCHER, S.; SCHULZ, A.; FRENZEL, L. P.; CLAASEN, J.; KUTSCH, N.; KRAUSE, G.; MAYR, C.; ROSENWALD, A.; PLASS, C.; SCHULTZE, J. L.; HALLEK, M.; WENDTNER, C. M. miRNA deregulation by epigenetic silencing disrupts suppression of the oncogene *PLAG1* in chronic lymphocytic leukemia. **Blood**, v. 114, p. 3255-3264, 2009.

PARK, S. D. E.; MAGEE, D. A.; MCGETTIGAN, P. A.; TEASDALE, M. D.;

EDWARDS, C. J.; LOHAN, A. J.; MURPHY, A.; BRAUD, M.; DONOGHUE, M. T.; LIU, Y.; CHAMBERLAIN, A. T.; RUE-ALBRECHT, K.; SCHROEDER, S.; SPILLANE, C.; TAI, S.; BRADLEY, D. G.; SONSTEGARD, T. S.; LOFTUS, B. J.; MACHUGH, D. E. Genome sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography and evolution of cattle. **Genome biology**, v. 16, 234, 2015.

PATZ, M.; PALLASCH, C. P.; WENDTNER, C.-M. Critical role of microRNAs in chronic lymphocytic leukemia: overexpression of the oncogene *PLAG1* by deregulated miRNAs. **Leukemia & lymphoma**, v. 51, p. 1379-1381, 2010.

PEREIRA, A. G. T.; UTSUNOMIYA, Y. T.; MILANESI, M.; TORRECILHA, R. B. P.; CARMO, A. S.; NEVES, H. H. R.; CARVALHEIRO, R.; AJMONE-MARSAN, P.; SONSTEGARD, T. S.; SÖLKNER, J.; CONTRERAS-CASTILLO, C. J.; GARCIA, J. F. Pleiotropic genes affecting carcass traits in *Bos indicus* (Nellore) cattle are modulators of growth. **PLOS ONE**, v. 11, e0158165, 2016.

PÉREZ-O'BRIEN, A. M.; HÖLLER, D.; BOISON, S. A.; MILANESI, M.; BOMBA, L.; UTSUNOMIYA, Y. T.; CARVALHEIRO, R.; NEVES, H. H. R.; SILVA, M. V. G. B.; VAN TASSELL, C. P.; SONSTEGARD, T. S.; MÉSZÁROS, G.; AJMONE-MARSAN, P.; GARCIA, J. F.; SÖLKNER, J. Low levels of taurine introgression in the current Brazilian Nelore and Gir indicine cattle populations. **Genetics Selection Evolution (GSE)**, v. 47, 31, 2015.

PORTO-NETO, L. R.; SONSTEGARD, T. S.; LIU, G. E.; BICKHART, D. M.; DA SILVA, M. V. B.; MACHADO, M. A.; UTSUNOMIYA, Y. T.; GARCIA, J. F.; GONDRO, C.; VAN TASSELL, C. P. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. **BMC Genomics**, v. 14, 876, 2013.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I. W.; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage

analyses. **American Journal of Human Genetics**, v. 81, p. 559-575, 2007.

QUEIMADO, L.; LOPES, C.; DU, F.; MARTINS, C.; BOWCOCK, A. M.; SOARES, J.; LOVETT, M. Pleomorphic adenoma gene 1 is expressed in cultured benign and malignant salivary gland tumor cells. **Laboratory Investigation**, v. 79, p. 583-589, 1999.

RANDHAWA, I. A. S.; KHATKAR, M. S.; THOMSON, P. C.; RAADSMA, H. W. Composite Selection Signals for Complex Traits Exemplified Through Bovine Stature Using Multibreed Cohorts of European and African *Bos taurus*. **Genes Genomes Genetics (G3)**, v. 5, p. 1391-1401, 2015.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.r-project.org/>. Acesso em: 4 set. 2017.

ROBINSON, J. T.; THORVALDSDÓTTIR, H.; WINCKLER, W.; GUTTMAN, M.; LANDER, E. S.; GETZ, G.; MESIROV, J. P. Integrative genomics viewer. **Nature Biotechnology**, v. 29, p. 24-26, 2011.

ROHLAND, N.; HOFREITER, M. Ancient DNA extraction from bones and teeth. **Nature Protocols**, v. 2, p. 1756-1762, 2007.

SAATCHI, M.; SCHNABEL, R. D.; TAYLOR, J. F.; GARRICK, D. J. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. **BMC Genomics**, v. 15, 442, 2014.

SABETI, P. C.; REICH, D. E.; HIGGINS, J. M.; LEVINE, H. Z. P.; 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.

STEPHENS, J. C.; REICH, D. E.; GOLDSTEIN, D. B.; SHIN, H. D.; SMITH, M. W.; CARRINGTON, M.; WINKLER, C.; HUTTLEY, G. A.; ALLIKMETS, R.; SCHRIML, L.; GERRARD, B.; MALASKY, M.; RAMOS, M. D.; MORLOT, S.; TZETIS, M.; ODDOUX, C.; DI GIOVINE, F. S.; NASIOULAS, G.; CHANDLER, D.; ASEEV, M.; HANSON, M.; KALAYDJIEVA, L.; GLAVAC, D.; GASPARINI, P.; KANAVAKIS, E.; CLAUSTRES, M.; KAMBOURIS, M.; OSTRER, H.; DUFF, G.; BARANOV, V.; SIBUL, H.; METSPALU, A.; GOLDMAN, D.; SCHMIDTKE, J.; ESTIVIL, X.; O'BRIEN, S. J.; DEAN, M. Dating the origin of the CCR5-delta-32 AIDS resistance allele by the coalescence of haplotypes. **American Journal of Human Genetics**, v. 62, p. 1507-1515, 1998.

STUCKY, B. J. Seqtrace: A graphical tool for rapidly processing DNA sequencing chromatograms. **Journal of Biomolecular Techniques**, v. 23, n. 3, p. 90-93, 2012.

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, p. 522-528, 2009.

THE BOVINE HAPMAP CONSORTIUM. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. **Science**, v. 324, p. 528-532, 2009.

THE INKSCAPE TEAM. **Inkscape v0.48.4-r9939**. Disponível em: <https://inkscape.org/en/>. Acesso em: 4 set. 2017.

THORVALDSDÓTTIR, H.; ROBINSON, J. T.; MESIROV, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. **Briefings in Bioinformatics**, v. 14, p. 178-192, 2013.

UTSUNOMIYA, A. T. H.; SANTOS, D. J. A.; BOISON, S. A.; UTSUNOMIYA, Y. T.; MILANESI, M.; BICKHART, D. M.; AJMONE-MARSAN, P.; SÖLKNER, J.; GARCIA, J. F.; DA FONSECA, R.; DA SILVA, M. V. G. B. Revealing misassembled segments in

the bovine reference genome by high resolution linkage disequilibrium scan. **BMC Genomics**, v. 17, 705, 2016a.

UTSUNOMIYA, Y. T.; BOMBA, L.; LUCENTE, G.; COLLI, L.; NEGRINI, R.; LENSTRA, J. A.; ERHARDT, G.; GARCIA, J. F.; AJMONE-MARSAN, P. Revisiting AFLP fingerprinting for an unbiased assessment of genetic structure and differentiation of taurine and zebu cattle. **BMC Genetics**, v. 15, 47, 2014a.

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.; 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**, v. 14, 52, 2013.

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.; SOÏLKNER, J.; PORTO-NETO, L. R.; SCHENKEL, F. S.; MCEWAN, J.; COLE, J. B.; DA SILVA, M. V. G. B.; VAN TASSELL, C. P.; SONSTEGARD, T. S.; GARCIA, J. F. Genome-wide mapping of loci explaining variance in scrotal circumference in Nellore cattle. **PLOS ONE**, v. 9, e88561, 2014b.

UTSUNOMIYA, Y. T.; MILANESI, M.; UTSUNOMIYA, A. T. H.; AJMONE-MARSAN, P.; GARCIA, J. F. GHap: an R package for genome-wide haplotyping. **Bioinformatics**, v. 32, p. 2861–2862, 2016b.

VAN DYCK, F.; DECLERCQ, J.; BRAEM, C. V.; VAN DE VEN, W. J. M. *PLAG1*, the prototype of the PLAG gene family: Versatility in tumour development (review). **International Journal of Oncology**, v. 30, p. 765-774, 2007.

VOZ, M. L.; AGTEN, N. S.; VAN DE VEN, W. J. M.; KAS, K. *PLAG1*, the main translocation target in pleomorphic adenoma of the salivary glands, is a positive

regulator of IGF-II. **Cancer Research**, v. 60, p. 106-113, 2000.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer International Publishing, 2016. 260 p.

ZAVAREZ, L. B.; UTSUNOMIYA, Y. T.; CARMO, A. S.; NEVES, H. H. R.; CARVALHEIRO, R.; FERENCAKOVIC, M.; PÉREZ-O'BRIEN, A. M.; CURIK, I.; COLE, J. B.; VAN TASSELL, C. P.; DA SILVA, M. V. G. B.; SONSTEGARD, T. S.; SÖLKNER, J.; GARCIA, J. F. Assessment of autozygosity in Nellore cows (*Bos indicus*) through high-density SNP genotypes. **Frontiers in Genetics**, v. 5, p. 1-8, 2015.

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**, v. 10, R42, 2009.

APPENDIX A - Documentation of the GHap package

1. Tutorial 1 - Importing phased data

Example input files can be created using the command:

```
# Copy the example data in the current working directory
library(GHap)
ghap.makefile()
```

The dataset comprises genotypes from the International HapMap Project Phase 3 (THE INTERNATIONAL HAPMAP 3 CONSORTIUM, 2010), which includes 1,011 subjects (from 11 populations) and 20,000 SNPs (randomly sampled from chromosome 2) mapped to the NCBI build 36 (hg18) assembly. The *ghap.loadphase()* function is responsible for loading phased chromosomes from an input file and converting them into a native **GHap.phase** object. A detailed description of this object can be found in the documentation of the function. To load the example data in the package we can run:

```
#Load haplotype object
phase <- ghap.loadphase(
  samples.file = "human.samples",
  markers.file = "human.markers",
  phase.file = "human.phase"
)
/    Reading in marker map information... Done.
/    A total of 20000 markers were found for chromosome 2.
/    Reading in sample information... Done.
/    A total of 1011 individuals were found in 11 populations.
/    Reading in phased genotypes... (may take a few minutes for large datasets)
/    Your GHap.phase object was successfully loaded without apparent errors.
```

The current version of the package only supports phased data of one chromosome at a time. However, once haplotypes have been called, multiple chromosomes can be loaded.

2. Tutorial 2 - Subsetting, exporting and merging phased objects

The *ghap.subsetphase()* function can take any combination of markers and individuals and subset the **GHap.phase** object. This is achieved by setting undesired markers and individuals to **FALSE**. Inactivated individuals and markers are then ignored by all other functions taking a **GHap.phase** object as input. For instance, we know that markers with low polymorphic information content may result in rare HapAlleles. If downstream analyses do not benefit from rare HapAlleles (e.g., haplotype association), it may be beneficial to prune these markers out prior to haplotyping. The code below shows how to subset markers with a minor allele frequency of at least 5%:

```
# Subset data - markers with maf > 0.05
maf <- ghap.maf(phase, ncores = 2)
markers <- phase$marker[maf > 0.05]
phase <- ghap.subsetphase(phase, unique(phase$id), markers)
/    Subsetting 1011 individuals and 17267 markers... Done.
/    Final data contains 1011 individuals and 17267 markers.
```

GHap.phase objects can also be exported to text files:

```
# Output data
ghap.outphase(phase, "example")
/     Preparing example.markers... Done.
/     Preparing example.samples... Done.
/     Preparing example.phase... Done.
```

It is also possible to merge two distinct **GHap.phase** objects with the

ghap.merge() function. There are three possible merging tasks: (i) Objects 1 and 2 have the same set of markers but different individuals; (ii) 2 - Objects 1 and 2 have different sets of markers (with potential overlaps) but the same individuals; (iii) Objects 1 and 2 have different sets of markers and individuals (with potential overlaps). Currently, GHap only supports task 1. This is because phase information may not derive from a consensus marker panel in task 2, and task 3 has the additional problem of forcing missing genotypes.

```
# Select ASW and CEU individuals
ASW.ids <- unique(phase$id[phase$pop=="ASW"])
CEU.ids <- unique(phase$id[phase$pop=="CEU"])

# Subset data
phase.ASW <- ghap.subsetphase(phase, ASW.ids, markers)
/    Subsetting 63 individuals and 17267 markers... Done.
/    Final data contains 63 individuals and 17267 markers.
phase.CEU <- ghap.subsetphase(phase, CEU.ids, markers)
/    Subsetting 117 individuals and 17267 markers... Done.
/    Final data contains 117 individuals and 17267 markers.

# Merge phase.ASW and phase.CEU
phase.merge <- ghap.mergephase(phase.ASW, phase.CEU)
/     Creating the new GHap.phase object... Done.
/    Your GHap.phase object was successfully merged without apparent errors.
```

3. Tutorial 3 - Haplotyping

In principle, the user can provide the coordinates of any arbitrary haplotype block (HapBlock). In GHap, we provide means to generate coordinates for HapBlocks based on sliding windows of markers. This strategy is particularly useful in genome-wide scans.

```
# Generate blocks of 5 markers sliding 5 markers at a time
blocks.mkr <- ghap.blockgen(phase, windowsize = 5, slide = 5, unit = "marker")

# Generate blocks of 100 kb sliding 100 kb at a time
blocks.kb <- ghap.blockgen(phase, windowsize = 100, slide = 100, unit = "kbp")
```

By default all blocks are constrained to a minimum of two markers. This behaviour can be adjusted by setting the *nsnp* argument to a different value. The extent of overlap between consecutive blocks can be controlled via the slide argument, depending on how fine the user wishes the genome-wide scan to be. Once HapBlocks have been defined, haplotype genotypes (HapGenotypes) can be determined:

```
# Generate matrix of haplotype genotypes
ghap.haplotyping(phase, blocks.mkr, batchsize = 100, ncores = 2, outfile = "human")
/   Processing 3453 blocks in:
/   1 batches of 53
/   34 batches of 100
/   3453 blocks written to file
```

By default all HapAlleles are included in the output. If intended, the user can exclude the minor HapAllele by setting the *drop.minor* argument to **TRUE**. Additionally, the *freq* argument allows for exclusion of HapAlleles outside of a specified frequency range. Control of memory usage and process parallelization is achieved through the arguments *batchsize* and *ncores*.

4. Tutorial 4 - Importing and manipulating haplotype data

After HapAlleles have been scored, the data can be loaded into R using the *ghap.loadhaplo()* function:

```
# Load haplotype genotypes
haplo <- ghap.loadphase(
   hapsamples.file = "human.hapsamples",
   hapalleles.file = "human.hapalleles",
   hapgenotypes.file = "human.hapgenotypes"
)
/   Reading in haplotype allele information... Done.
/   A total of 60002 haplotype alleles were found.
/   Reading in sample information... Done.
/   A total of 1000 individuals were found in 1 populations.
/   Reading in haplotype genotypes... (may take a few minutes for large datasets)
/   Your GHap.haplo object was successfully loaded without apparent errors.
```

Similar to the **GHap.phase** object, the user can also subset, merge and export **GHap.haplo** objects. For instance:

```
# Randomly select 500 individuals
ids <- sample(x = haplo$id, size = 500, replace = FALSE)

# Subset data
haplo.sub <- ghap.subsethaplo(haplo,ids,haplo$allele.in)
/   Subsetting 500 individuals and 60002 haplotype alleles... Done.
/   Final data contains 500 individuals and 60002 haplotype alleles.

# Output new GHap.haplo object
ghap.outhaplo(haplo = haplo.sub, outfile = "humansub")
/   Preparing humansub.hapsamples... Done.
/   Preparing humansub.hapalleles... Done.
/   Preparing humansub.hapgenotypes... Done.
```

5. Tutorial 5 - Haplotype statistics

For each HapAllele, the *ghap.hapstats()* function retrieves absolute and relative frequencies, expected and observed number of homozygotes, and different tests for deficit of homozygotes in comparison to Hardy-Weinberg Equilibrium (HWE) expectations.

```
hapstats <- ghap.hapstats(haplo, ncores = 2)
str(hapstats)
/      'data.frame':   60002 obs. of  14 variables:
/      $ BLOCK    : chr  "CHR2_B1" "CHR2_B1" "CHR2_B1" "CHR2_B1" ...
/      $ CHR      : chr  "2" "2" "2" "2" ...
/      $ BP1      : num  18228 18228 18228 18228 18228 ...
/      $ BP2      : num  75360 75360 75360 75360 75360 ...
/      $ ALLELE   : chr  "ATAGT" "ATAAC" "ATGGC" "GGAAC" ...
/      $ N        : num  2 4 5 10 42 ...
/      $ FREQ     : num  0.000989 0.001978 0.002473 0.004946 0.020772 ...
/      $ O.HOM    : num  0 0 0 0 0 1 14 17 14 524 ...
/      $ O.HET    : num  2 4 5 10 42 56 123 142 170 328 ...
/      $ E.HOM    : num  0.000989 0.003956 0.006182 0.024728 0.436202 ...
/      $ RATIO    : num  1 1 1.01 1.02 1.44 ...
/      $ BIN.logP : num  0.00043 0.00172 0.00268 0.01074 0.18948 ...
/      $ POI.logP : num  0.00043 0.00172 0.00268 0.01074 0.18944 ...
/      $ TYPE     : chr  "MINOR" "REGULAR" "REGULAR" "REGULAR" ...
```

The function also assigns a TYPE category to each HapAllele:

“ABSENT” = the frequency of the allele is 0;
“SINGLETON” = unique haplotype of its block with frequency 1 (i.e., monomorphic block);
“MINOR” = the least frequent haplotype of its block (in the case of ties, only the first haplotype is marked);
“MAJOR” = the most frequent hapotype of its block (ties are also resolved by marking the first haplotype);
“REGULAR” = the haplotype does not fall into any of the previous categories.

Categories “SINGLETON”, “MINOR” and “MAJOR” only apply to blocks where frequencies sum to 1. The *ghap.blockstats()* function summarizes HapAllele statistics per block and retrieves the expected heterozygosity and the number of alleles per HapBlock. For instance:

```
blockstats <- ghap.blockstats(hapstats, ncores = 2)
head(blockstats,n=2)
/          BLOCK  CHR     BP1     BP2       EXP.H  N.ALLELES
/  01 CHR2_B1      2   18228   75360  0.5128683          10
/  11 CHR2_B2      2   90190  109437  0.7139595          15
```

Notice that calculation of expected heterozygosity will not be reliable when HapAlleles are prunned out by frequency during haplotyping. Therefore, the function will return **NA** for blocks where HapAllele frequencies do not sum to unity. Also, when the dataset contains multiple populations the expected heterozygosity and the number of alleles will be very high.

6. Tutorial 6 - Relationship matrix and PCA

The example below computes a kinship matrix from HapGenotypes and plots the first two eigenvectors of a principal components analysis of this matrix. Notice that absent, singleton and minor alleles should be excluded from computations.

```
# Subset major and regular alleles
haplo <- ghap.subsethaplo(haplo, haplo$id,
                                hapstats$TYPE %in% c("REGULAR","MAJOR"))
/    Subsetting 1011 individuals and 56572 haplotype alleles... Done.
/    Final data contains 1011 individuals and 56572 haplotype alleles.

# Compute Kinship matrix
K <- ghap.kinship(haplo, batchsize = 100)
/    Processing 56572 HapAlleles in 566 batches.
/    Inactive alleles will be ignored.
/    Preparing 1011 x 1011 kinship matrix.
/    56572 HapAlleles processed.

# PCA analysis
pca <- ghap.pca(haplo,K)

# Plot
plot(x=pca$eigenvec$PC1, y=pca$eigenvec$PC2, xlab="PC1", ylab="PC2", pch="")
pop <- pca$eigenvec$POP
pop.col <- as.numeric(as.factor(pop))
pop <- sort(unique(pop))
legend("bottomleft", legend = pop, col = 1:length(pop), pch = 1:length(pop), ncol = 3)
points(x=pca$eigenvec$PC1, y=pca$eigenvec$PC2,
        pch = pop.col, col = pop.col, cex = 1.2)
```

7. Tutorial 7 - Haplotype divergence analysis

The example below compares the CEU and CHB populations for HapBlocks on chromosome 2:

```
# Compute haplotype allele statistics for each group
haplo <- ghap.subsethaplo(haplo,haplo$id,rep(TRUE,times=haplo$nalleles))
CHB.ids <- haplo$id[which(haplo$pop=="CHB")]
CEU.ids <- haplo$id[which(haplo$pop=="CEU")]
haplo <- ghap.subsethaplo(haplo,CHB.ids,haplo$allele.in)
CHB.hapstats <- ghap.hapstats(haplo,ncores = 2)
haplo <- ghap.subsethaplo(haplo,CEU.ids,haplo$allele.in)
CEU.hapstats <- ghap.hapstats(haplo,ncores = 2)
haplo <- ghap.subsethaplo(haplo,c(CHB.ids,CEU.ids),haplo$allele.in)
TOT.hapstats <- ghap.hapstats(haplo,ncores = 2)
haplo <- ghap.subsethaplo(haplo,haplo$id,rep(TRUE,times=haplo$nalleles))

# Compute haplotype block statistics for each group
CHB.blockstats <- ghap.blockstats(CHB.hapstats, ncores = 2)
CEU.blockstats <- ghap.blockstats(CEU.hapstats, ncores = 2)
TOT.blockstats <- ghap.blockstats(TOT.hapstats, ncores = 2)

# Calculate Fst
fst<-ghap.fst(CHB.blockstats, CEU.blockstats, TOT.blockstats)

# Plot results
top.fst <- fst[fst$FST == max(fst$FST, na.rm=TRUE),]
plot(x = (fst$BP1+fst$BP2)/2e+6, y = fst$FST, pch = "",
      ylab = expression(paste("Haplotype ", F[ST])),
      xlab = "Chromosome 2 (in Mb)", ylim=c(0,1)
)
abline(v=108.7, col="gray")
```

```
points(x = (fst$BP1+fst$BP2)/2e+6, y = fst$FST, pch = 20, col="#471FAA99")
points(x = (top.fst$BP1+top.fst$BP2)/2e+6, y = top.fst$FST, pch = 20, col="red")
text(x = 125, y = max(fst$FST, na.rm=TRUE), "EDAR", col="red")
```

Ideally, similar to the case of HapAllele and HapBlock statistics, the F_{ST} analysis should be carried out on the full set of HapAlleles, rather than a frequency-prunned subset.

8. Tutorial 8 - Haplotype ancestry

GHap offers a way to calculate the probability that a given HapAllele from a tested population was inherited from one of the tested parental populations. For instance, using CEU and YRI as proxy parental populations for ASW, we could assign HapAlleles in ASW to CEU or YRI using the following code:

```
# Compute haplotype allele statistics for each group
haplo <- ghap.subsethaplo(haplo,haplo$id,rep(TRUE,times=haplo$nalleles))
ASW.ids <- unique(haplo$id[haplo$pop=="ASW"])
YRI.ids <- unique(haplo$id[haplo$pop=="YRI"])
CEU.ids <- unique(haplo$id[haplo$pop=="CEU"])
haplo <- ghap.subsethaplo(haplo,YRI.ids,haplo$allele.in)
YRI.hapstats <- ghap.hapstats(haplo,ncores = 2)
haplo <- ghap.subsethaplo(haplo,CEU.ids,haplo$allele.in)
CEU.hapstats <- ghap.hapstats(haplo,ncores = 2)
haplo <- ghap.subsethaplo(haplo,ASW.ids,haplo$allele.in)
ASW.hapstats <- ghap.hapstats(haplo,ncores = 2)
haplo <- ghap.subsethaplo(haplo,haplo$id,rep(TRUE,times=haplo$nalleles))
```

```
# Find haplotype origin
# ASW is the test population. YRI and CEU are used as parental populations
# The frequency threshold is set to 0.05 and the probability of assignment to 0.60
ancestry <- ghap.ancestral(ASW.hapstats, YRI.hapstats, CEU.hapstats, 0.05, 0.60)
ancestry <- ancestry[ancestry$FREQ.TEST > 0,]
str(ancestry)
/      'data.frame':   38561 obs. of  11 variables:
/     $ BLOCK             : chr  "CHR2_B1" "CHR2_B1" "CHR2_B1" "CHR2_B1" ...
/     $ CHR               : chr  "2" "2" "2" "2" ...
/     $ BP1               : num  18228 18228 18228 18228 18228 ...
/     $ BP2               : num  75360 75360 75360 75360 75360 ...
/     $ ALLELE            : chr  "ATAAC" "ATAGC" "GGAAC" "GGAGC" ...
/     $ FREQ.TEST         : num  0.00794 0.05556 0.01587 0.18254 0.07143 ...
/     $ FREQ.PARENT1 : num  0 0.087 0 0.1435 0.0783 ...
/     $ FREQ.PARENT2 : num  0 0.00855 0 0 0 ...
/     $ PROB.PARENT1 : num  0 0.911 0 1 1 ...
/     $ PROB.PARENT2 : num  0 0 0 0 0 ...
/     $ ORIGIN            : chr  "UNK" "PARENT1" "UNK" "PARENT1" ...
```

9. Tutorial 9 - Linear mixed model analysis

GHap implements a wrapper of the lme4 package (BATES et al., 2015) to fit generalized linear mixed models of the form:

$$g(\mu_{y|u}) = \mathbf{Xb} + \mathbf{Zu}$$

where $g(\ .\ )$ is a link function, $\mu_{y|u}$ is the expectation of phenotypes conditional on random effects, **b** is a vector of unobserved fixed effects, **X** is a matrix relating phenotypes to **b**, **u** is a vector of random effects $\sim N(\mathbf{0}, \mathbf{K}\sigma_u^2)$, and **Z** is an incidence matrix relating phenotypes to **u**. Random effects can be partitioned into subgroups with different covariance matrices. For instance, if we let **K** be the HapAllele

relationship matrix, then **u** becomes the HapAllele-based polygenic effects/breeding values, and σ_u^2 becomes the variance due to HapAlleles. Importantly, any arbitrary **K** matrix is admitted, such that one may fit models combining pedigree and haplotype relationships (e.g., single-step GWAS analysis, see Wang et al., 2012).

In the example below we simulate a quantitative trait in Europeans with 50% heritability, where two major HapAlleles account for 50% of the genetic variance. Repeated records are taken for each individual. However, the dataset is unbalanced, such that subjects can have between 0 and 30 measurements.

```
# Subset common haplotypes in Europeans
EUR.ids <- haplo$id[haplo$pop %in% c("TSI","CEU")]
haplo <- ghap.subsethaplo(haplo,EUR.ids,rep(TRUE,times=haplo$nalleles))
hapstats <- ghap.hapstats(haplo, ncores = 2)
common <- hapstats$TYPE %in% c("REGULAR","MAJOR") &
hapstats$FREQ > 0.05 &
hapstats$FREQ < 0.95
haplo <- ghap.subsethaplo(haplo,EUR.ids,common)

#Compute relationship matrix
K <- ghap.kinship(haplo, batchsize = 100)

# Quantitative trait with 50% heritability
# Unbalanced repeated measurements (0 to 30)
# Two major haplotypes accounting for 50% of the genetic variance
myseed <- 123456789
set.seed(myseed)
major <- sample(which(haplo$allele.in == TRUE),size = 2)
g2 <- runif(n = 2, min = 0, max = 1)
g2 <- (g2/sum(g2))*0.5
sim <- ghap.simpheno(haplo, kinship = K, h2 = 0.5, g2 = g2, nrep = 30,
                              balanced = FALSE, major = major, seed = myseed)
```

```
#Fit model using REML
model <- ghap.lmm(fixed = phenotype ~ 1, random = ~ individual,
                  covmat = list(individual = K), data = sim$data)

#Estimated heritability and repeatability
model$vcp/sum(model$vcp)

#True versus estimated breeding values
plot(model$random$individual,sim$u,xlab="Estimated BV",ylab="True BV")
abline(0,1)
summary(lm(sim$u ~ as.numeric(model$random$individual)))
```

10. Tutorial 10 - Association analysis

The *ghap.assoc()* function regresses a response variable on one HapAllele at a time, treating HapAlleles as fixed effects. The example below takes the simulated data from the previous tutorial and regresses residuals and genomic estimated breeding values onto HapAlleles.

```
#HapAllele GWAS using GEBVs as response
pheno <- model$random$individual
gwas1 <- ghap.assoc(response = pheno, haplo = haplo, ncores = 4)

#HapAllele GWAS using GEBVs as response
#Weight observations by number of repeated measurements
pheno <- model$random$individual
w <- table(sim$data$individual)
w <- w + mean(w)
w <- w[names(pheno)]
gwas2 <- ghap.assoc(response = pheno, haplo = haplo, ncores = 4, weights = w)
```

```
#HapAllele GWAS using residuals as response
pheno <- model$residuals
names(pheno) <- sim$data$individual
gwas3 <- ghap.assoc(response = pheno, haplo = haplo, ncores = 4)

#Plot results
plot(gwas1$BP1/1e+6,gwas1$logP,pch=20,col="darkgreen",ylim=c(0,20),
      xlab="Position (in Mb)",ylab=expression(-log[10](p)))
points(gwas2$BP1/1e+6,gwas2$logP,pch=20,col="gray")
points(gwas3$BP1/1e+6,gwas3$logP,pch=20,col="blue")
abline(v=haplo$bp1[major]/1e+6,lty=3)
abline(h=-log10(0.05/nrow(gwas1)),lty=3)
legend("topleft",legend = c("GEBVs","weighted GEBVs","residuals"),
          pch = 20,col=c("darkgreen","gray","blue"))
```

11. Tutorial 11 - BLUP of haplotypes

HapAlleles can also be treated as random effects with the *ghap.blup()* function. Random effects can be iteratively updated through the haploweights argument following the single-step GWAS approach (WANG et al., 2012):

```
#BLUP GWAS
gebvs <- model$random$individual
gebvsw <- table(sim$data$individual)
gebvsw <- gebvsw + mean(gebvsw)
gebvsw <- gebvsw[names(gebvs)]
Kinv <- ghap.kinv(K)
gwas.blup <- ghap.blup(gebvs = gebvs, haplo = haplo, gebvsweights = gebvsw,
                                ncores = 4, invcov = Kinv)
plot(gwas.blup$BP1/1e+6,gwas.blup$pVAR*100,pch=20,
      xlab="Position (in Mb)",ylab="Variance explained (%)")
```

```
#BLUP with one update
w <- gwas.blup$VAR*nrow(gwas.blup)
K2 <- ghap.kinship(haplo=haplo,weights = w)
Kinv2 <- ghap.kinv(K2)
gwas.blup2 <- ghap.blup(gebvs = gebvs, haplo = haplo, invcov = Kinv2, ncores = 2,
                        gebvsweights = gebvsw, haploweights = w)
plot(gwas.blup2$BP1/1e+6,gwas.blup2$pVAR*100,pch=20,
     xlab="Position (in Mb)",ylab="Variance explained (%)")
abline(v=haplo$bp1[major]/1e+6)
```

12. Tutorial 12 - Haplotype profiling

The profile for each individual is calculated as:

$$\sum_{i=1}^{m} h_i a_i$$

where relative to HapAllele *i*, h_i is the number of copies and a_i is a user-defined score. By default, if scores are provided for only a subset of the HapAlleles, the missing alleles scores will be set to zero. This function has the same spirit as the profiling routine implemented in the score option in PLINK (PURCELL et al., 2007; CHANG et al., 2015). This function can be useful for analyses involving cross-validation of genomic predictions based on BLUP solutions of HapAllele effects or scoring admixture proportions from the output of *ghap.ancestral()*. Below is an example using simulated scores from a normal distribution:

```
# Create a score data.frame
score <- NULL
score$BLOCK <- haplo$block
score$CHR <- haplo$chr
score$BP1 <- haplo$bp1
score$BP2 <- haplo$bp2
score$ALLELE <- haplo$allele
set.seed(1988)
score$SCORE <- rnorm(length(score$ALLELE))
score <- data.frame(score,stringsAsFactors = FALSE)
score$CENTER <- 0
score$SCALE <- 1

# Compute profiles
profile <- ghap.profile(score, haplo, ncores = 2)
head(profile)
/        POP          ID       PROFILE
/     1 ASW   NA19904   -38.410381
/     2 ASW   NA20340   -12.250027
/     3 ASW   NA20297   -45.473774
/     4 ASW   NA20281    -7.360974
/     5 ASW   NA20348   -36.271198
/     6 ASW   NA20300  40.912226
```

13. Methods 1 - Format

The supported format is composed of three files with suffix:

.samples = space-delimited file without header containing two columns: Population and ID. Please notice that the Population column serves solely for the purpose of grouping samples, so the user can define any arbitrary family/cluster/subgroup and use as a “population” tag.

.markers = space-delimited file without header containing five columns: Chromosome, Marker, Position (in bp), Reference Allele (A0) and Alternative Allele (A1). Markers should be on a single chromosome and sorted by position.

.phase = space-delimited file without header containing the phased genotype matrix. The dimension of the matrix is expected to be m x 2n, where m is the number of markers and n is the number of individuals. Alleles must be coded as 0 and 1. No missing values are allowed.

See below an example of five individuals from the ASW population with phased genotypes for five markers on chromosome 2:

.samples file	.markers file	.phase file
ASW NA19904	2 rs13383216 18228 A G	1 1 1 1 1 1 1 1 1 1
ASW NA20340	2 rs13386087 24503 G T	0 0 0 0 0 0 0 0 0 0
ASW NA20297	2 rs10179984 33092 A G	1 0 1 0 0 0 0 0 1 1
ASW NA20281	2 rs300761 60074 A G	0 1 0 0 1 1 0 1 0 1
ASW NA20348	2 rs6749571 72820 C G	0 0 0 0 0 0 0 1 0 0

This format is conveniently obtained with very little manipulation from the output of widely used phasing software, such as SHAPEIT2 (O’CONNELL et al., 2014). For instance, to format your SHAPEIT2 files with UNIX standard commands use:

```
tail -n +3 shapeit2_file.sample | cut -d' ' -f1,2 > GHapfile.samples
cut -d' ' -f1-5 shapeit2_file.haps > GHapfile.markers
cut -d' ' -f1-5 --complement shapeit2_file.haps > GHapfile.phase
```

14. Methods 2 – Haplotyping algorithm

Let a haplotype library (HapLibrary) be the collection of observed HapAlleles for a given HapBlock. The haplotyping procedure implemented in GHap is straightforward: each HapAllele in the library is treated as a pseudo-marker, and HapGenotypes are scored as 0, 1 or 2 HapAllele copies. Using the example data from the last section, assume the user wishes to call haplotypes for the first three markers. The algorithm works as follows: First, we crop the matrix at the selected markers (for the sake of clarity, we will transpose the matrix and represent subjects in rows and markers in columns):

POP	ID	rs13383216	rs13386087	rs10179984
ASW	NA19904	1	0	1
ASW	NA19904	1	0	0
ASW	NA20340	1	0	1
ASW	NA20340	1	0	0
ASW	NA20297	1	0	0
ASW	NA20297	1	0	0
ASW	NA20281	1	0	0
ASW	NA20281	1	0	0
ASW	NA20348	1	0	1
ASW	NA20348	1	0	1

The HapLibrary is created based on the unique HapAlleles:

HapAllele1: 101 (GGG)
HapAllele2: 100 (GGA)

Then, for each HapAllele, individual HapGenotypes are scored based on the number of copies:

POP	ID	GGG	GGA
ASW	NA19904	1	1
ASW	NA20340	1	1
ASW	NA20297	0	2
ASW	NA20281	0	2
ASW	NA20348	2	0

The procedure is then repeated for each HapBlock. The haplotyping function outputs three files with suffix:

.hapsamples = space-delimited file without header containing two columns: Population and Individual ID.

.hapalleles = space-delimited file without header containing five columns: Block Name, Chromosome, Start and End Position (in bp), and Haplotype Allele.

.hapgenotypes = space-delimited file without header containing the haplotype genotype matrix (coded as 0, 1 or 2 copies of the haplotype allele). The dimension of the matrix is *m* x *n*, where *m* is the number of haplotype alleles and *n* is the number of subjects.

For instance, the example above would yield the following data:

.hapsamples file	.hapalleles file	.hapggenotypes file
ASW NA19904	CHR2_B1 2 18228 33092 GGG	1 1 0 0 2
ASW NA20340	CHR2_B1 2 18228 33092 GGA	1 1 2 2 0
ASW NA20297		
ASW NA20281		
ASW NA20348		

15. Methods 3 - Haplotype statistics

Relative to HapAllele *i*, let p_i, h_i and n represent the relative frequency, the number of homozygotes, and the number of subjects, respectively. Also, let S_i be some test statistic or score for the HapAllele, representing the goodness-of-fit of h_i to HWE expectations. The *ghap.hapstats()* function computes three candidate methods for S_i:

Method 1. The number of homozygotes for haplotype *i* is expected to be $E[h_i] = np_i^2$ under HWE. Provided we observed $O[h_i]$ homozygotes, deviations from HWE expectations can be expressed in terms of the expected-to-observed ratio:

$$S_i = \frac{E[h_i] + \alpha_1}{O[h_i] + \alpha_2}$$

where α_1 and α_2 are shrinkage parameters. The purpose of the shrinkage parameters is to regularize the scores towards a ratio of α_1 / α_2, being particularly useful in cases where the number of observed homozygotes is close to zero. As the null ratio value is 1 (i.e., expected and observed counts are equal), a reasonable choice of shrinkage parameters is $\alpha_1 = \alpha_2 = 1$ (the default in GHap), which in practice introduces a bias equivalent to that of one additional expected and one additional observed homozygote. For a more detailed review on shrinkage expected-to-observed (or observed-to-expected) ratio, see Norén, Hopstadius and Bate (2013).

Method 2. Under the null hypothesis of HWE, $h_i \sim \text{Binomial}(n, p_i^2)$, with $E[h_i] = np_i^2$ and $VAR[hi] = np_i^2(1 - p_i^2)$. Therefore, the probability of observing h_i or less homozygotes given the haplotype is in HWE is:

$$Pr(X \leqslant h_i) = \sum_{j \leqslant h_i} \binom{n}{j} p_i^{2j} (1 - p_i^2)^{n-j}$$

where *X* is a random draw from the Binomial distribution.

Method 3. Provided n is large, $h_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = E[h_i] = VAR[h_i] = np_i^2$. This leads to probability:

$$Pr(X \leqslant h_i) = e^{-\lambda_i} \sum_{j \leqslant h_i} \frac{\lambda_i^j}{j!}$$

Note that the variance in the Binomial model is smaller than in the Poisson model, which in practice results in more conservative probabilities in the latter case.

16. Methods 4 - Haplotype coding for regression and relationship matrix

Consider a multi-allelic locus and let alleles 1, 2, . . . , h be ordered with frequencies $\mathbf{p} = [p_1 \;\; p_2 \;\; \ldots \;\; p_h]'$ (from lowest to highest). Following Falconer (1960), the genotypic value associated with genotype ij can be decomposed into:

$$g_{ij} = \mu + u_{ij} + \delta_{ij}$$

where μ, u_{ij} and δ_{ij} are the genotypic mean, the breeding value (BV) and the dominance deviation, respectively. Here we will focus only on the BV, such that the dominance deviation will be treated as a residual effect. Assuming Hardy-Weinberg Equilibrium (HWE), the BV can be partitioned into allelic effects (Da, 2015):

$$u_i = \sum_{j \neq i} p_i \alpha_{ij}$$

where is the average effect of substituting allele i by allele j. It follows that $\alpha_{ii} = 0$ and $\alpha_{ij} = -\alpha_{ji}$, such that there are only $h - 1$ independent substitution effects to consider, which can be expressed as the effects of replacing a reference allele by any other in the same locus. Da (2015) proposed setting the most frequent allele as the reference. However, since the choice is arbitrary and do not affect the resulting BV, we will consider at first the least frequent allele (i.e., allele 1) as the reference instead for later convenience. In this setting, the BV can be expressed as:

$$u_{ij}=\sum_{k=2}^{h} m_{ij,k}\alpha_{1k}$$

where $m_{ij,k}$ is a scalar taking values:

$-(0 - 2p_k)$, for $i, j \neq k$
$-(1 - 2p_k)$, for $i \neq j$ but $i = k$ or $j = k$
$-(2 - 2p_k)$, for $i = j = k$

So far all substitution effects α_{1k} are expressed in the direction of allele 1. However, we wish to derive substitution effects in the direction of each allele by treating them as the reference, and use allele 1 as the basis for contrasts. Since we established that $\alpha_{1k} = -\alpha_{1k}$, we can re-write the BV as:

$$u_{ij}=\sum_{k=2}^{h} -m_{ij,k}\alpha_{k1}$$

where $-m_{ij,k}$ is a scalar taking values:

$(0 - 2p_k)$, for 0 copies of allele k
$(1 - 2p_k)$, for 1 copy of allele k
$(2 - 2p_k)$, for 2 copies of allele k

Since the $2p_k$ term represents the mean allele count when HWE is assumed, an alternative coding not requiring HWE is obtained from replacing $2p_k$ by the sample mean. This is the approach we adopted in GHap. If the locus is bi-allelic, the allele coding collapses to the genotype coding used for SNP markers. In fact, SNP-based regression is revealed here as a special case of haplotype-based regression, where HapBlocks are bi-allelic and of size 1 bp. This coding also reveals that regression on HapAlleles is in fact equivalent to fitting haplotypes as pseudo bi-allelic markers, provided that an arbitrary HapAllele (in this case the minor HapAllele) has been discarded (i.e., set as the basis for contrasts). Without loss of generality, rare and

nearly fixed HapAlleles can also be discarded in order to reduce the number of predictors, procedure that is analogous to exclusion of SNPs by minor allele frequency in SNP-based regression.

The coding presented above is also used to compute the haplotype-based relationship matrix. Briefly, let **M** be the centered *N* x *H* matrix of HapGenotypes, where *N* is the number of observations and *H* is the number of HapAlleles. The HapAllele correlations among individuals can be computed as:

$$\mathbf{K} = q\mathbf{MDM}'$$

where $\mathbf{D} = \text{Diag}(d_i)$, d_i is the weight of HapAllele *i* (default $d_i = 1$) and $q = \text{tr}(\mathbf{MDM}')^{-1}N$. Notice that this is a generalization of the SNP-based genomic relationship matrix (VANRADEN, 2008).

17. Methods 5 - Regression treating haplotypes as fixed effects

The least squares regression procedure in Ghap tests each HapAllele at a time for association with phenotypes. The fixed effect, error variance and test statistic of a given HapAllele are estimated as:

$$\alpha = (\mathbf{m}'\mathbf{m})^{-1}\mathbf{m}'\mathbf{y}$$
$$\text{VAR}[\alpha] = (\mathbf{m}'\mathbf{m})^{-1}\sigma_e^2$$
$$t^2 = \alpha^2 / \text{VAR}[\alpha]$$

Under the null hypothesis that the regression coefficient is zero $t^2 \sim \chi^2(\nu = 1)$. Although nothing prevents the user to fit raw phenotypes, the use of adjusted records accounting for covariates, polygenic effects and other potential random effects is advisable. For instance, residuals from the mixed model analysis could be used as the response variable for regression on HapAlleles. The user must be aware of two known caveats associated with this approach: (i) by pre-adjusting records instead of estimating HapAllele effects based on generalized least squares equations we ignore covariance structure and therefore bias the estimates downwards (SVISHCHEVA et

al., 2012); (ii) each HapAllele being tested is also included in the kinship matrix, such that the HapAllele is included twice in the model: as fixed and random effect. This problem is known as proximal contamination (LISTGARTEN et al., 2012).

In the first case, we can use genomic control to recover p-values to an unbiased scale (DEVLIN; ROEDER, 1999; AMIN; VAN DUIJN; AULCHENKO, 2007). However, not much can be done regarding the estimates of the effects. As a general recommendation, if the user is only interested in p-values, the regression analysis discussed here should be sufficient. When effect estimates are of interest, the user can select genome-wide significant HapAlleles and include them as fixed effects in the full mixed model. For the second case, a leave-one-chromosome-out (LOCO analysis) procedure can mitigate proximal contamination (YANG et al., 2014). An alternative to these methods is to use polygenic effects as response instead of residuals. However, this can lead to a higher false-positive rate (EKINE et al., 2014).

18. Methods 6 - Regression treating haplotypes as random effects

Recall that the generalized linear mixed model assumes:

$$\mathbf{u} \mid \sigma_u^2 \sim N(0, \mathbf{K}\sigma_u^2)$$

If we let $\mathbf{K} = q\mathbf{MDM}'$, it follows that $\mathbf{u} = \mathbf{M}\alpha$. This means that we can convert between individual breeding values and HapAllele effects (STRANDÉN; GARRICK, 2009):

$$\alpha = q\mathbf{DM}'\mathbf{K}^{-1}\mathbf{u}$$

19. Methods 7 - Fixation index

Haplotype-based F_{ST} analyses are supported by the *ghap.fst()* function. Calculations are based on the multi-allelic formula (NEI, 1973):

$$F_{ST} = (H_T - H_S)/H_T$$

where H_T is the total gene diversity (i.e., expected heterozygosity in the population) and H_S is the sub-population gene diversity (i.e., the average expected heterozygosity in the sub-populations).

20. Methods 8 - Ancestry assignment

The procedure follows the method described by Bolormaa et al. (2011):

$$Pr(parent1) = p_{parent1} / (p_{parent1} + p_{parent2})$$
$$Pr(parent1) = p_{parent2} / (p_{parent1} + p_{parent2})$$

where $p_{parent1}$ and $p_{parent2}$ are the HapAllele frequencies in the first and second parental populations, respectively. Assignments are performed as follows: if the probability of one of the parents exceeds a user-defined threshold (default = 0.60), the HapAllele origin is assigned to that parental population. Parental probabilities are set to zero if the HapAllele frequency in the parental populations are lower than a certain threshold (default = 0.05).

21. Methods 9 - Using GHap outputs in third-party software

When the haplotyping procedure is performed using very large datasets, post hoc analyses may be too computationally demanding to be performed in R. Also, existing pipelines designed to analyze bi-allelic SNP data can be extended to the analysis of haplotypes by simply incorporating the output generated by the *ghap.hap2tped()* function in GHap. This function creates a set of files that mimic a standard PLINK (PURCELL et al., 2007; CHANG et al., 2015) *tped* file, where HapAllele counts 0, 1 and 2 are recoded as NN, NH and HH genotypes (N = NULL and H = haplotype allele), as if HapAlleles were bi-alelic markers. This coding scheme is acceptable for any given analysis relying on genotype counts, as long as the user specifies that the analysis should be done using the H allele as reference for counts. You can specify reference alleles using the **.tref** file in PLINK with the reference-allele command. The name for each pseudo-marker is composed by a

concatenation (separated by "_") of block name, start, end, and haplotype allele identity. Pseudo-marker positions are computed as (start+end)/2. Of note, for applications such as GWAS it is advisible to output only MAJOR and REGULAR HapAlleles, since SINGLETONS and MINOR HapAlleles will not contribute to the analysis.

The following lines of code show one example of how the output from GHap can be articulated with analyses that are routinely applied to unphased SNP marker data. First, we can export the **GHap.haplo** object to use in PLINK:

```
# Subset common haplotypes
hapstats <- ghap.hapstats(haplo, ncores = 2)
common <- hapstats$TYPE %in% c("REGULAR","MAJOR") &
                hapstats$FREQ > 0.05 &
                hapstats$FREQ < 0.95
haplo <- ghap.subsethaplo(haplo,unique(haplo$id),common)

# Output GHap.haplo object
ghap.outhaplo(haplo = haplo, outfile = "humansub")

# Convert to tped
ghap.hap2tped(infile = "humansub", outfile = "humansub")
```

Then, we can use PLINK to perform a principal components analysis on our data:

```
#Converting the tped output to PLINK binary
plink --tfile humansub --reference-allele humansub.tref --make-bed --out humansub

#Performing PCA analysis in PLINK
#Correlations and scale with the GHap package are almost perfect (r = 0.999)
plink --bfile humansub --reference-allele humansub.tref --pca 2 --out humansub
```

22. Methods 10 - Handling multiple chromosomes and analysis of single marker data

By default GHap works at single chromosome data, specially when it comes to the haplotyping procedure. However, once HapAlleles have been called on each chromosome, the user can choose to load one chromosome at a time or to load all chromosomes together with *ghap.loadhaplo()*. This can be typically achieved by the use of the *ghap.mergehaplo()* function, or alternatively by concatenation of single chromosome files. You can also fool GHap to take in single SNP data (say you wish to compare haplotype x single SNP association results). To do so, you only need to see SNPs as distinct 1bp HapBlocks, and count number of copies of a particular allele (e.g., minor or reference allele). Then, *ghap.loadhaplo()* will naturally load single SNP data. Be aware that for some analyses special consideration may be required, so be sure that you know exactly what you are doing!

23. Benchmarking

Benchmarking of the main tasks in the package was first performed in a Dell PowerEdge-T410 workstation with 16 GB RAM and two 64-bit Intel Xeon 2.13 GHz CPUs, running R v3.2.5 under Ubuntu 10.04 LTS. Performance was evaluated using the HapMap data with varying number of cores.

Table 1. Benchmarking of GHap with varying numbers of cores. A total of 1,011 samples and 20,000 markers were used. Time was measured in seconds and averaged over 10 replicates.

	Number of cores			
Task	1	2	4	8
Load data	9.86 ± 0.11	-	-	-
Filtering*	8.20 ± 0.23	4.20 ± 0.14	2.60 ± 0.13	1.59 ± 0.04
Haplotyping	169.50 ± 1.01	82.47 ± 0.22	42.00 ± 0.33	23.00 ± 0.19

*Minor allele frequency prunning and subsetting

Another benchmarking was conducted to assess the influence of dataset size on performance. This benchmarking was done using a Dell T5500 workstation with 24 GB and 64-bit Intel Xeon 3.07GHz CPU, running R v3.2.3 under Red Hat Enterprise Linux Workstation release 6.7.

Table 2. Benchmarking of GHap with 8 cores and varying numbers of markers and subjects. Time was measured in seconds and averaged over 10 replicates.

Markers	Samples	Task		
		Load data	Filtering	Haplotyping
10,000	1,000	2.62 ± 0.01	0.34 ± 0.04	6.29 ± 0.10
	5,000	13.14 ± 0.03	0.91 ± 0.01	28.33 ± 0.25
	10,000	26.04 ± 0.07	1.49 ± 0.05	57.42 ± 0.41
	50,000	134.44 ± 0.54	7.98 ± 0.20	282.44 ± 1.43
100,000	1,000	28.02 ± 0.07	3.30 ± 0.09	76.64 ± 0.31
	5,000	143.01 ± 0.22	9.95 ± 0.13	286.89 ± 0.90
	10,000	281.83 ± 0.91	26.73 ± 0.51	561.97 ± 1.48
	50,000*	-	-	-

*The analysis of 50,000 samples and 100,000 markers consumed more than the maximum RAM available (24GB) and was unfeasible using available hardware

In summary, Ghap scales linearly as a function of markers or subjects. We noticed a limitation in the analysis of a large number of individuals, but this was related to RAM availability. Analyses of such large datasets may be accomplished using high-performance computing facilities or by subdividing the data in batches with smaller sample sizes.

24. References

AMIN, N.; VAN DUIJN, C. M.; AULCHENKO, Y. S. A Genomic Background Based Method for Association Analysis in Related Individuals. **PLOS ONE**, v. 2, e1274, 2007.

BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting Linear Mixed-Effects Models Using lme4. **Journal of Statistical Software**, v. 67, p. 1-48, 2015.

BOLORMAA, S.; HAYES, B. J.; HAWKEN, R. J.; ZHANG, Y.; REVERTER, A.; GODDARD, ME. Detection of chromosome segments of zebu and taurine origin and their effect on beef production and growth. **Journal of Animal Science**. v. 89, p.2050-2060, 2011.

CHANG, C. C.; CHOW, C. C.; TELLIER, L. C.; VATTIKUTI, S.; PURCELL, S. M.; LEE, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. **GigaScience**, v. 4, 7, 2015.

DA, Y. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. **BMC Genetics**, v. 16, 144, 2015.

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

EKINE, C. C.; ROWE, S. J.; BISHOP, S. C.; DE KONING D. J. Why breeding values estimated using familial data should not be used for genome-wide association studies. **Genes Genomes Genetics (G3)**, v. 4, p. 341-347, 2014.

LISTGARTEN, J.; LIPPERT, C.; KADIE, C. M.; DAVIDSON, R. I.; ESKIN, E.; HECKERMAN, D. Improved linear mixed models for genome-wide association studies. **Nature Methods**, v. 9, p. 525-526, 2012.

NEI, M. Analysis of Gene Diversity in Subdivided Populations. **Proceedings of the National Academy of Sciences of the United States of America (PNAS USA)**, v. 70, p. 3321-3323, 1973.

NORÉN, G.N.; HOPSTADIUS, J.; BATE, A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Statistical Methods in Medical Research, v. 22, p. 57-69, 2013.

O'CONNELL, J.; GURDASANI, D.; DELANEAU, O.; PIRASTU, N.; ULIVI, S.; COCCA, M.; TRAGLIA, M.; HUANG, J.; HUFFMAN, J. E.; RUDAN, I.; MCQUILLAN, R.; FRASER, R. M.; CAMPBELL, H.; POLASEK, O.; ASIKI, G.; EKORU, K.; HAYWARD, C.; WRIGHT, A. F.; VITART, V.; NAVARRO, P.; ZAGURY, J. F.; WILSON, J. F.; TONIOLO, D.; GASPARINI, P.; SORANZO, N.; SANDHU, M. S.; MARCHINI, J. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. **PLOS Genetics**, v. 10, e1004234, 2014.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I. W.; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, v. 81, p. 559-575, 2007.

STRANDÉN, I.; GARRICK, D. J. Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. **Journal of Dairy Science**, v. 92, p. 2971-2975, 2009.

SVISHCHEVA, G. R.; AXENOVICH, T. I.; BELONOGOVA, N. M.; VAN DUIJN, C. M.; AULCHENKO, Y. S. Rapid variance components-based method for whole-genome association analysis. Nature Genetics, v. 44, p. 1166-1170, 2012.

THE INTERNATIONAL HAPMAP 3 CONSORTIUM. Integrating common and rare genetic variation in diverse human populations. **Nature**, v. 467, 52-58, 2010.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**. v. 91, p. 4414-4423, 2008.

WANG, H.; MISZTAL, I.; AGUILAR, I.; LEGARRA, A.; MUIR, W. M. Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, v. 94, p. 73-83, 2012.

YANG, J.; ZAITLEN, N. A.; GODDARD, M. E.; VISSCHER, P. M.; PRICE, A. L. Advantages and pitfalls in the application of mixed-model association methods. **Nature Genetics**, v. 46, p. 100-106, 2014.

APPENDIX B - Supplementary Analyses

1. Analysis of Y chromosome haplotypes supports introgression of *Q* in Japanese cattle

The HapMap data are summarized in Table 1. The Illumina® BovineHD panel included 1,224 Y-linked markers. Ideally, genotypes at these markers should be recalled manually since both the standard cluster file and the GenTrain2 algorithm in GenomeStudio assume diploidy. However, as this process is extremely laborious, we opted to analyze only markers with obvious two-states hemizygous genotypes in the HapMap data. First, we retained only markers with GenTrain Score greater than 0.80. This threshold guaranteed that clusters AA, AB and BB were well-shaped and clearly separated. Second, animals with call rate lower than 95% at Y-linked markers were excluded, which mainly caused females to be removed from downstream analyses. Next, all remaining heterozygous calls were set to missing and homozygous genotypes were replaced by hemizygous states (0 for AA and 1 for BB). Lastly, only markers with call rate equal to 100% and minor allele frequency greater than 5% in the overall breed data were preserved. This filtering procedure resulted in 29 markers and 563 individuals, which yielded five different haplotypes (YHapG1 – YHapG5, Table 2).

YHapG1 and YHapG2 were identified as *B. indicus* haplotypes, whereas YHapG3, YHapG4 and YHapG5 were recognized as *B. taurus* haplotypes. Additionally, YHapG3 was nearly fixed in six out of the nine Northwestern European breeds that were highly selected for *Q*. Importantly, similar to crossbred and admixed populations, Wagyu presented multiple Y chromosome haplotypes, including YHapG3, which is consistent with admixture and supports the hypothesis of introgression of *Q* from Northwestern European cattle into Japanese breeds.

Table 1. Breed information for the Bovine HapMap data.

Species	Breed	Origin	Longitude[a]	Latitude[a]	Purpose	Sample size	Weight (in kg)[b]	Withers height (in cm)[b]
Bos taurus	Blonde d'Aquitaine	France	0.0878906	43.6821749	Beef	5	1300	165
	Normande	France	0.1712529	48.8798704	Dairy	5	1100	155
	Norwegian Red	Norway	8.4689460	60.4720240	Dairy	17	1000	142
	Red Angus	Britain	-2.0942780	57.1497170	Beef	11	1000	145
	Holstein	Netherlands	5.7817542	53.1641642	Dairy	71	1100	165
	Angus	Britain	-2.0942780	57.1497170	Beef	47	1000	145
	Limousin	France	1.5696018	45.8932231	Beef	51	1050	144
	Charolais	France	4.2752780	46.4344400	Beef	40	1150	146
	Senepol	Caribbean (composite)	-64.8347992	17.7245968	Beef	12	930	136
	Wagyu	Japan	138.2529240	36.2048240	Beef	13	940	142
	Hereford	Britain	-2.6544182	52.0765164	Beef	38	1050	135
	Guernsey	Britain	-2.5852780	49.4656910	Dairy	21	760	140
	Simmental (Fleckvieh)	Switzerland	8.2275120	46.8181880	Dairy	10	1200	152
	Piedmontese	Italy	7.5153885	45.0522366	Beef	24	950	150
	Montbéliard	France	6.0000000	47.5833300	Dairy	5	1100	148
	Jersey	Jersey	-2.1312500	49.2144390	Dairy	46	600	127
	Romagnola	Italy	9.4702148	45.2377644	Beef	34	700	145
	Brown Swiss (Braunvieh)	Switzerland	10.0000000	46.8812215	Dairy	24	1050	152
	N'Dama	Guinea	-9.6966450	9.9455870	Dairy	24	370	116
	Lagunaire	Benin	2.3158340	9.3076900	Beef	5	200	100
Bos indicus	Nellore	India	79.9864560	14.4425987	Beef	35	570	150
	Gyr	India	71.1923805	22.2586520	Dairy	30	544	140
Bos taurus x Bos indicus	Beefmaster	United States (composite)	-99.9018131	31.9685988	Beef	24	1202	NA
	Santa Gertrudis	United States (composite)	-97.8561090	27.5158689	Beef	35	861	NA
	Brahman	United States (composite)	-81.1637245	33.8360810	Beef	49	887	126
	Brangus	United States (composite)	-91.6634483	29.9110378	Beef	13	900	135
	Sheko	Ethiopia (ancient crossbred)	40.4896730	9.1450000	Dairy	18	NA	NA
Bubalus bubalis	Buffalo	Southern Asia	71.1923805	22.2586520	Dual	7	NA	NA
Bos gaurus	Gaur	Southern Asia	76.4563087	25.0376400	Dual	2	NA	NA
Bos grunniens	Yak	Southern Asia	91.1175250	29.6475350	Dual	2	NA	NA

[a]Longitude and latitude data were approximated based on the geographical origin of each breed.

[b]Average male body weight and withers height data were obtained from the Domestic Animal Diversity Information System of the Food and Agricultural Organization (FAO) of the United Nations. Available at: http://dad.fao.org/. Accession date: 10 Nov 2016.

Table 2. Frequency of Y chromosome haplotypes[a] in the HapMap data.

Breed	YHapG1: GTCTCCCGCCGGATAG GGCGCAGCCCACT	YHapG2: GTCTCCTACCGGATAG GGCGCACCCCACT	YHapG3: ACTCGGTATTACGCAA AATATGCTTCTAC	YHapG4: ACTCGGTATTACGCTA GATATGCTTCAAC	YHapG5: ACTCGGTATTACGCAA GATATGCTTTAAC
Blonde d'Aquitaine	0.0000	0.0000	0.0000	1.0000	0.0000
Normande	0.0000	0.0000	1.0000	0.0000	0.0000
Norwegian Red	0.0000	0.0000	1.0000	0.0000	0.0000
Red Angus	0.0000	0.0000	1.0000	0.0000	0.0000
Holstein	0.0000	0.0000	1.0000	0.0000	0.0000
Angus	0.0000	0.0000	1.0000	0.0000	0.0000
Limousin	0.0000	0.0000	0.0000	1.0000	0.0000
Charolais	0.0000	0.0000	0.0000	1.0000	0.0000
Senepol	0.0000	0.0000	0.1667	0.8333	0.0000
Wagyu	0.0000	0.0000	0.1538	0.5385	0.3077
Hereford	0.0000	0.0000	0.7576	0.0606	0.1818
Guernsey	0.0000	0.0000	0.0000	1.0000	0.0000
Simmental (Fleckvieh)	0.0000	0.0000	0.2000	0.8000	0.0000
Piedmontese	0.0000	0.0000	0.0000	0.6250	0.3750
Montbéliard	0.0000	0.0000	0.0000	1.0000	0.0000
Jersey	0.0000	0.0000	0.1395	0.0000	0.8605
Romagnola	0.0000	0.0000	0.0000	1.0000	0.0000
Brown Swiss (Braunvieh)	0.0000	0.0000	0.0000	1.0000	0.0000
N'Dama	0.0000	0.0000	0.0000	1.0000	0.0000
Nellore	1.0000	0.0000	0.0000	0.0000	0.0000
Gyr	1.0000	0.0000	0.0000	0.0000	0.0000
Beefmaster	0.3043	0.0000	0.6957	0.0000	0.0000
Santa Gertrudis	0.9688	0.0000	0.0000	0.0313	0.0000
Brahman	1.0000	0.0000	0.0000	0.0000	0.0000
Brangus	0.0000	0.0000	0.8571	0.1429	0.0000
Sheko	0.6000	0.4000	0.0000	0.0000	0.0000

[a]Haplotypes were defined based on a subset of 29 polymorphic markers mapping to chromosome Y.

2. Coalescence at the *POLLED* locus implies parallel selection for *Q* and polledness

Two candidate mutations causing polledness in *B. taurus* have been previously identified on chromosome 1 (CHR1) (MEDUGORAC et al., 2012; ALLAIS-BONNET et al., 2013): an ~80 kbp duplication of Friesian origin (P_F, CHR1:1909352-1989480) and an 212 bp duplication of Celtic origin (P_C, CHR1:1705834-1706045). Similarly to *Q*, both mutations appear to occur in a Northwestern European selective sweep spanning CHR1:1693164-2018403 . Also, the earliest predicted carrier of *Q* in our Nellore data was polled, raising the hypothesis of concomitant spread of polledness and *Q* in worldwide cattle. The main caveat in the analysis of the *POLLED* locus is that both short-horned and polled breeds present loss of diversity with similar haplotypes in this CHR1 region (ALLAIS-BONNET et al., 2013), indicating that a putative selection event for decreased horn size preceded the occurrence of both P_C or P_F and that other variants in the region might also contribute to horn size and morphology. In this case, identical-by-state haplotypes constructed from array data can be in linkage with P_C, P_F or even with the short-horn allele, preventing reliable estimation of identity-by-descent. Nevertheless, it was still possible to use haplotype analyses to estimate coalescence of the selection signatures for decreased horn size and polledness in European cattle and detect *B. taurus* introgression at this locus in the *B. indicus* lineage.

First, in order to confirm whether the polled allele in Nellore was indeed of *B. taurus* origin, we regressed horned/polled phenotypes onto CHR1 haplotypes. Prior to phasing, misassembled segments on CHR1 were excluded, following our previous report (UTSUNOMIYA et al., 2016). Associations mapped to a ~559 kbp segment on CHR1:1473797-2032837 ($p = 2.28 \times 10^{-28}$), which included the relevant signature region (Figure 1a). To simplify the analysis, we focused on the six-markers haplotypes overlapping with the positions of P_C (CHR1:1702350-1721777) and P_F (CHR1:1905048-1925652). At the P_C and P_F regions, the significant haplotypes were TGAAAG and TGTAGG, respectively. Both were the major haplotypes in most of the European breeds in the HapMap data, except for Romagnola (see Tables 3 and 4). These haplotypes were also rare or absent in African *B. taurus*, *B. indicus* and

outgroup species. This confirms that polledness in Nellore is a *B. taurus* contribution. Frequency of the *PLAG1* haplotype GGGTTC was highly correlated with frequency of TGAAAG or TGTAGG (r = 0.787, p = 2.53 x 10^{-7}) in worldwide cattle (Table 5). Although the frequencies of GGGTTC (0.1778) and TGAAAG/TGTAGG (0.1007) were similar in our sample of Nellore bulls, GGGTTC occurrence did not differ between carriers (0.224) and non-carriers (0.181) of TGAAAG/TGTAGG. Since *POLLED* and *PLAG1* were located in different chromosomes, independent assortment of these haplotypes may have favored an even spread of *Q* between polled and horned Nellore after introgression. This is plausible given selection for polledness is currently practiced by specific groups of breeders.

Considering all Northwestern European breeds together (Figure 1b), time to coalescence of the two *POLLED* haplotypes was estimated at ~2274 yBP (95% CI [2070, 2485]), which implies selection for decreased horn size since the Iron Ages. However, when only polled breeds carrying the Celtic allele were considered (i.e., Angus, Red Angus and Simmental), coalescence was much younger, indicating a selective sweep approximately ~464 yBP (95% CI [370, 560]). Considering the Friesian allele (i.e., Holstein and Jersey), the signature dated to ~565 yBP (95% CI [445, 650]). These estimates were close to the coalescence for the *Q* haplotype. Although we still cannot directly connect the origin of the *Q* mutation to the origin of P_C/P_F, altogether our results provide evidence that increased body size and polledness were selected in parallel in Europe. This implies that polled European cattle are likely to carry *Q*, and that either early carriers were polled or that polledness first appeared in populations selected for *Q*.

The estimate in Nellore carriers was ~203 yBP (95% CI [145, 270]), suggesting that introgression of TGAAAG/TGTAGG occurred earlier than *Q*, most likely as soon as the first imported animals arrived from India to Brazil. Of note, due to the cryptic allelic heterogeneity at the *POLLED* haplotypes, we could not link this time estimate with the actual introgression of the polled phenotype into Nellore.

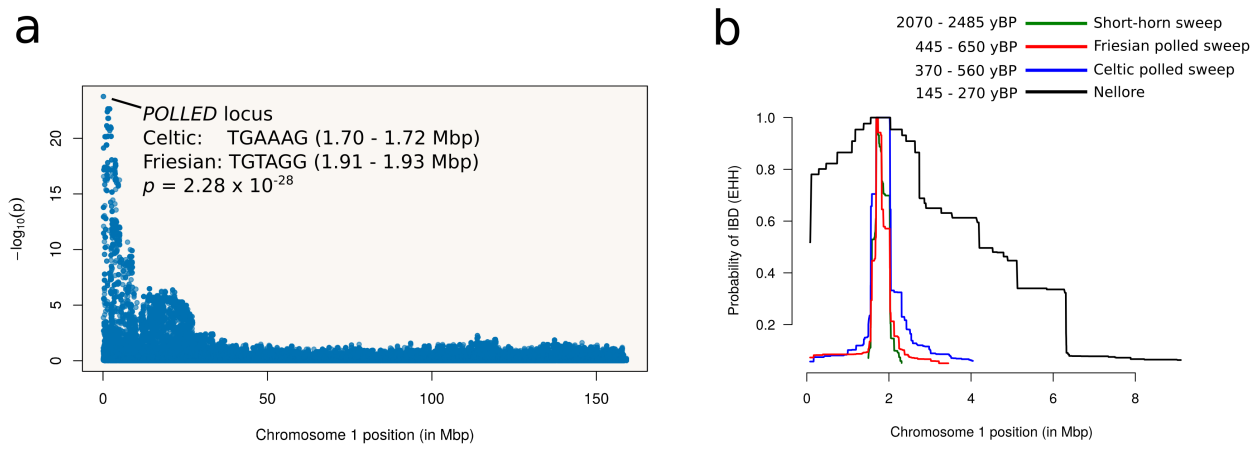

Figure 1. Association mapping and coalescence for the *POLLED* locus. (a) Scatterplot showing maximum association (p = 2.28 x 10^{-28}) for horned/polled phenotypes in Nellore cattle mapping to the European *POLLED* selective sweep. The significant haplotypes are of *B. taurus* origin. (b) Extended haplotype homozygosity (EHH). Simultaneous analysis of all Northwestern European breeds reveals that selection for decreased horn size dates back to the Iron Ages. The Celtic and Friesian signatures are much more recent and coalesce to a period close to the selection for *Q*. The polled haplotypes are shown to have been introgressed into Nellore cattle during early imports to Brazil.

Table 3. Haplotype frequencies at the Celtic *POLLED* locus in the Bovine HapMap data.

Breed	CAGGGA	CAGGAG	CAGAGG	CAGAAG	CGGAGA	CGAAGA	CGAAGG	CGAAAG	TAGGGA	TAGGAG	TAGAGA	TAGAGG	TAGAAG	TGAGGA	TGAAGA	TGAAGG	TGAAAA	TGAAAG
Blonde d'Aquitaine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Normande	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	0.0000	0.0000	0.0000	0.0000	0.0385	0.0385	0.0000	0.8846
Norwegian Red	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1875	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8125
Red Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Holstein	0.0000	0.0833	0.0000	0.0208	0.0000	0.0833	0.0000	0.0000	0.0000	0.0625	0.0000	0.0000	0.0000	0.0000	0.0417	0.0000	0.0000	0.7083
Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Limousin	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9615
Charolais	0.0000	0.0000	0.0000	0.0000	0.0000	0.0395	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0132	0.0000	0.0000	0.9474
Senepol	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Wagyu	0.0000	0.0000	0.0000	0.0000	0.0000	0.0735	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0294	0.0294	0.2206	0.3824	0.0000	0.2647
Hereford	0.0000	0.0000	0.0000	0.0000	0.0000	0.0238	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9762
Guernsey	0.0375	0.0000	0.0000	0.0000	0.0000	0.0125	0.0000	0.0500	0.0000	0.0000	0.0000	0.0000	0.0125	0.0000	0.0000	0.0375	0.0000	0.8500
Simmental	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Piedmontese	0.0000	0.0000	0.1250	0.0000	0.0208	0.0417	0.0000	0.0833	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0625	0.0000	0.6667
Montbéliard	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1000	0.0000	0.0000	0.9000
Jersey	0.0000	0.0000	0.0000	0.0098	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0196	0.0000	0.0000	0.9706
Romagnola	0.0143	0.0000	0.0143	0.0000	0.0143	0.1429	0.0143	0.4286	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1143	0.0000	0.2571
Brown Swiss	0.0000	0.0000	0.0000	0.0000	0.0000	0.1957	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0109	0.0000	0.7935
N'Dama	0.0833	0.0000	0.3167	0.0000	0.0500	0.0833	0.1333	0.0000	0.0667	0.0000	0.0167	0.0167	0.0000	0.0000	0.0000	0.1833	0.0000	0.0500
Lagunaire	0.0408	0.0000	0.0714	0.0000	0.0816	0.2143	0.1633	0.0204	0.0510	0.0000	0.0000	0.0510	0.0000	0.0000	0.0102	0.0918	0.0000	0.2041
Nellore	0.0000	0.0000	0.2571	0.0000	0.1429	0.0143	0.3286	0.0000	0.1571	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1000	0.0000	0.0000
Gyr	0.3333	0.0000	0.0000	0.0278	0.0000	0.1667	0.1667	0.0000	0.0833	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0833	0.0000	0.1389
Beefmaster	0.0070	0.0000	0.0000	0.0000	0.0000	0.0211	0.0000	0.0070	0.0000	0.0000	0.0000	0.0000	0.0352	0.0000	0.0141	0.0000	0.0000	0.9155
Santa Gertrudis	0.0294	0.0000	0.0000	0.0000	0.0000	0.0294	0.0000	0.0294	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0294	0.0000	0.8824
Brahman	0.2292	0.0000	0.0000	0.0000	0.0000	0.2083	0.2708	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0625	0.2292
Brangus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Sheko	0.0000	0.0000	0.0000	0.0000	0.0000	0.4000	0.4000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2000
Buffalo	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0714	0.0000	0.0000	0.0000	0.0000	0.8571	0.0000	0.0000	0.0000	0.0714	0.0000	0.0000
Gaur	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Yak	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Haplotypes were constructed based on Illumina® BovineHD data of markers BovineHD0100000544 (rs109832358), BovineHD0100000545 (rs135487276), BovineHD0100000546 (rs132949427), BovineHD0100000547 (rs134388759), BovineHD0100000548 (rs109756112) and BovineHD0100000549 (rs133176033).

Table 4. Haplotype frequencies at the Friesian *POLLED* locus in the Bovine HapMap data.

Breed	CACCAA	CACAGG	CGCCAA	CGCCGG	CGTCGG	CGTAGG	TATCGG	TATAGG	TGCCAG	TGCCGA	TGCCGG	TGTCGG	TGTAGA	TGTAGG
Blonde d'Aquitaine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Normande	0.0000	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8000
Norwegian Red	0.0588	0.0000	0.1176	0.0000	0.0294	0.0000	0.0000	0.0294	0.0000	0.0000	0.0000	0.0000	0.0000	0.7647
Red Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Holstein	0.0000	0.0000	0.2887	0.0000	0.0211	0.0282	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6620
Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Limousin	0.0000	0.0000	0.0980	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9020
Charolais	0.0000	0.0000	0.0750	0.0000	0.0250	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0125	0.0000	0.8875
Senepol	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Wagyu	0.0000	0.0000	0.5000	0.0000	0.0385	0.0385	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	0.0000	0.3846
Hereford	0.0395	0.0000	0.0132	0.0000	0.0000	0.0000	0.0000	0.0132	0.0000	0.0000	0.0000	0.0000	0.0000	0.9342
Guernsey	0.0238	0.0000	0.1667	0.0000	0.0000	0.0238	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7857
Simmental (Fleckvieh)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Piedmontese	0.2083	0.0000	0.0833	0.0000	0.1042	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6042
Montbéliard	0.0000	0.0000	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8000
Jersey	0.0543	0.0000	0.0109	0.0000	0.0109	0.0000	0.0000	0.0109	0.0000	0.0000	0.0000	0.0000	0.0000	0.9130
Romagnola	0.0000	0.0000	0.3088	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3382	0.0000	0.3529
Brown Swiss (Braunvieh)	0.0000	0.0000	0.1042	0.0000	0.1875	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0417	0.0000	0.6667
N'Dama	0.1042	0.0000	0.5208	0.0000	0.0000	0.0000	0.3542	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0208
Lagunaire	0.0000	0.0000	0.1000	0.0000	0.0000	0.0000	0.7000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2000
Nellore	0.0714	0.0000	0.2143	0.0000	0.0000	0.0000	0.0000	0.0000	0.0143	0.0857	0.0571	0.5571	0.0000	0.0000
Gyr	0.1167	0.0000	0.1167	0.0000	0.0000	0.0000	0.0000	0.0000	0.1167	0.0000	0.0333	0.5667	0.0000	0.0500
Beefmaster	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1250	0.0000	0.0000	0.0625	0.0000	0.8125
Santa Gertrudis	0.0143	0.0000	0.0143	0.0000	0.1143	0.0000	0.0143	0.0143	0.0143	0.0000	0.0143	0.0000	0.0143	0.7857
Brahman	0.1327	0.0000	0.2041	0.0000	0.0102	0.0000	0.0000	0.0000	0.0306	0.0102	0.0000	0.3878	0.0000	0.2245
Brangus	0.0000	0.0000	0.0385	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9615
Sheko	0.0000	0.0000	0.3611	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0833	0.4167	0.0000	0.1389
Buffalo	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8571	0.0000	0.1429	0.0000	0.0000
Gaur	0.0000	0.0000	0.0000	0.2500	0.0000	0.0000	0.0000	0.0000	0.0000	0.7500	0.0000	0.0000	0.0000	0.0000
Yak	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000

Haplotypes were constructed based on Illumina® BovineHD data of markers BovineHD0100000595 (rs136319130), BovineHD0100000596 (rs133617837), BovineHD0100000597 (rs134380776), BovineHD0100000598 (rs137606204), ARS-BFGL-NGS-76349 (rs109797076) and BovineHD0100000599 (rs133424654).

Table 5. Haplotype diversity at the CHR14:24973324-25012733 *PLAG1* locus in the Bovine HapMap data

Breed	ATACCT	ATATCT	ATATTC	AGATCC	AGATTC	AGGTTC	GGATTC	GGGTTC (Q)
Blonde d'Aquitaine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Normande	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Norwegian Red	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Red Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Holstein	0.0070	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9930
Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0106	0.9894
Limousin	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0196	0.9804
Charolais	0.0000	0.0000	0.0000	0.0000	0.0250	0.0000	0.0000	0.9750
Senepol	0.0417	0.0000	0.0000	0.0000	0.0833	0.0000	0.0000	0.8750
Wagyu	0.0000	0.0000	0.0000	0.0000	0.1538	0.0000	0.0000	0.8462
Hereford	0.0000	0.0000	0.0000	0.0000	0.0790	0.0000	0.0790	0.8420
Guernsey	0.0000	0.0000	0.0000	0.0000	0.1670	0.0000	0.1190	0.7140
Simmental (Fleckvieh)	0.0000	0.0000	0.0000	0.0000	0.3500	0.0000	0.1500	0.5000
Piedmontese	0.0000	0.0000	0.0000	0.0000	0.1880	0.0000	0.3330	0.4790
Montbéliard	0.0000	0.0000	0.0000	0.0000	0.4000	0.0000	0.2000	0.4000
Jersey	0.0000	0.0000	0.0000	0.0000	0.3910	0.0000	0.2500	0.3590
Romagnola	0.0000	0.0290	0.0000	0.2940	0.4410	0.0000	0.0590	0.1760
Brown Swiss (Braunvieh)	0.0000	0.0000	0.0000	0.0210	0.2920	0.3960	0.1670	0.1250
N'Dama	0.0000	0.0000	0.0000	0.0000	0.9170	0.0420	0.0210	0.0210
Lagunaire	0.0000	0.0000	0.0000	0.0000	0.3000	0.0000	0.7000	0.0000
Nellore	0.8000	0.1140	0.0140	0.0000	0.0140	0.0000	0.0000	0.0570
Gyr	0.7170	0.2170	0.0330	0.0000	0.0330	0.0000	0.0000	0.0000
Beefmaster	0.1040	0.0000	0.0000	0.0000	0.0630	0.0000	0.0000	0.8330
Santa Gertrudis	0.0430	0.0000	0.0000	0.0000	0.0140	0.0000	0.0140	0.9290
Brahman	0.2960	0.0310	0.0000	0.0000	0.0200	0.0000	0.0000	0.6530
Brangus	0.0000	0.0380	0.0000	0.0000	0.0000	0.0000	0.0000	0.9620
Sheko	0.3890	0.1670	0.0830	0.0000	0.2500	0.0830	0.0000	0.0280
Buffalo	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gaur	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Yak	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Haplotypes were constructed based on Illumina® BovineHD data of markers BovineHD1400007249 (rs110243083), BovineHD1400007250 (rs136888475), BovineHD1400007253 (rs109636480), BovineHD1400007254 (rs135404594), BovineHD1400007257 (rs134286310) and BovineHD1400007258 (rs135538206).

3. Assessment of genotype imputation based on a reduced set of 24 bulls

We evaluated whether the 24 sequenced Nellore bulls could serve as a reference set for imputation of HD genotypes up to sequence variants. The main idea was that, if accurate predictions of whole genome sequence genotypes could be achieved using these bulls, the haplotype association analysis presented in the main paper could be replaced by direct analysis of sequence variants. Since test animals had only HD data available, extraction of HD genotypes was performed on the 24 sequenced bulls, and test animals had their genotypes reduced to panels of smaller marker density that partially overlapped with the HD array (Table 6). These panels included the industry's standard Illumina® BovineSNP50 v2 (50K) and two *B. indicus*-specific panels, namely GeneSeek® Genomic Profiler Bos Indicus HD (GGP75Ki) and Illumina® Z-Chip v1 (ZChip). FImpute v2.2 (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014) was used to impute genotypes, as it has been shown to yield comparable accuracies with competing algorithms with substantially smaller run times (PICOLLI et al., 2014; VENTURA et al., 2014). In comparison to Beagle (BROWNING; BROWNING, 2008), FImpute was found to yield higher accuracies in a wide range of imputation scenarios in this Nellore population (CARVALHEIRO et al., 2014). Briefly, FImpute's method is based on a deterministic approach that assumes all animals are related to each other to some degree. Initially, genotypes are predicted using chromosome windows with shared long haplotypes (identical-by-descent segments due to recent ancestors). Then, genotypes are further imputed using shorter windows to capture information from more distant relatives. Accuracy of imputation was measured as the percentage of correctly imputed genotypes. Average accuracies were 81.50 ± 4.21%, 89.10 ± 2.73% and 82.41 ± 4.12% for 50K, GGP75Ki and ZChip, respectively. These results indicated that a larger sample of reference animals would be required for accurate imputation.

Table 6. Number of markers per assay used in the imputation analysis.

Assay name	Abbreviation	Manifest version	Genome assembly	Total number of markers	Markers in common with HD	Markers in common with HD after filter [a]
Illumina® BovineSNP50 v2	50K	C	UMDv3.1	54,609	49,915	21,239
GeneSeek® Genomic Profiler Bos Indicus HD	GGP75Ki	A	UMDv3.1	74,677	73,959	64.206
Illumina® Z-Chip v1	ZChip	A	UMDv3.1	27,533	27,202	21,333
Illumina® BovineHD	HD	A	UMDv3.1	786,799	786,799	447,617

[a]Autosomal markers that had call rate > 95% and minor allele frequency > 5%.

4. Influence of haplotype size and population structure on association analyses

We re-analyzed the birth weight data in order to assess how population stratification and haplotype size could affect the association results. We first used the same method described in the material and methods with haplotypes of size 1 (equivalent to performing a standard single-marker analysis), 5, 10 and 20, paralleling our original analysis of 6-markers haplotypes. Then, we repeated all analyses considering the following mixed linear model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{e}$$

where **y** is the vector of dEBVs, **1** is a vector of 1s, μ is the intercept, **u** is the vector of breeding values or polygenic effects (i.e., sum of the random effects of genome-wide markers), and **e** is the vector of residual effects. The model assumed $\mathbf{y} \sim N(\mathbf{1}\mu, \mathbf{V})$ for $\mathbf{V} = \mathbf{K}\sigma_u^2 + \mathbf{W}\sigma_e^2$, where **K** is the genomic relationship matrix (GRM) excluding the tested chromosome, **W** is the dEBV weight matrix as described in the material and methods, and σ_u^2 and σ_e^2 are the variance components related to breeding values and residual effects, respectively. Variance components were estimated using restricted maximum likelihood (REML) with the Newton-Raphson algorithm. Similarly to the original analysis, we computed $\alpha_k = (\mathbf{z}_k\mathbf{V}^{-1}\mathbf{z}_k)^{-1}\mathbf{z}_k\mathbf{V}^{-1}(\mathbf{y} - \mathbf{1}\mu)$ and $SE(\mathbf{a}_k) = VAR(\mathbf{a}_k)^{1/2} = (\mathbf{z}_k\mathbf{V}^{-1}\mathbf{z}_k)^{-1/2}$ for each haplotype *k*, and the association mapping was based on the two-tailed t-test α_k / $SE(\alpha_k)$. This approach was very similar to the standard Leave-One-Chromosome-Out Mixed Linear Model Association (MLMA-LOCO) method (YANG et al., 2014), except that our analysis considered haplotypes instead of single-markers, and that heterogeneity of residual variance was explicitly modeled.

We found that association results were robust in respect to variations in haplotype size and presence of population structure (Figure 2). Considering no correction for genetic stratification, the most significant associations were found at positions 26007360 (p = 4.63 x 10^{-18}), 24973324-25012733 (p = 4.77 x 10^{-17}), 25692794-25753013 (p = 6.35 x 10^{-17}) and 24406302-24478336 (p = 2.44 x 10^{-16}) for the analyses of haplotypes built over 1, 5, 10 and 20 markers, respectively. In the

case of the mixed model analysis, peaking positions were identified at 24473841 (p = 4.52 x 10^{-9}), 24472819-24478336 (p = 7.18 x 10^{-9}), 25692794-25753013 (p = 1.91 x 10^{-8}) and 24571130-24653754 (p = 4.08 x 10^{-8}) for windows of 1, 5, 10 and 20 markers, respectively. Altogether, these analyses suggested that the causal mutation is most likely located between positions 24406302 and 26007360 bp on chormosome 14. This ~1.6 Mbp window contained the fine-mapped region reported by Boitard et al. (2016) spanning 24.80 – 25.08 Mbp on CHR14, as well as the candidate quantitative trait nucleotides (QTNs) located between 24.97 and 25.05 Mbp reported by Karim et al. (2011) (Table 7). Therefore, we conclude that the ~39.5 kbp haplotype spanning positions 24973324-25012733 bp was a suitable predictor for the *Q* mutation, regardless of the existence of actual physical overlap between the two. As justified in CHAPTER 3, inclusion of SNP rs109815800 further improved tagging of *Q* in worldwide cattle breeds, which yielded a final ~42.3 kbp haplotype ranging from 24973324 to 25015640 bp on CHR14 (Table 8)

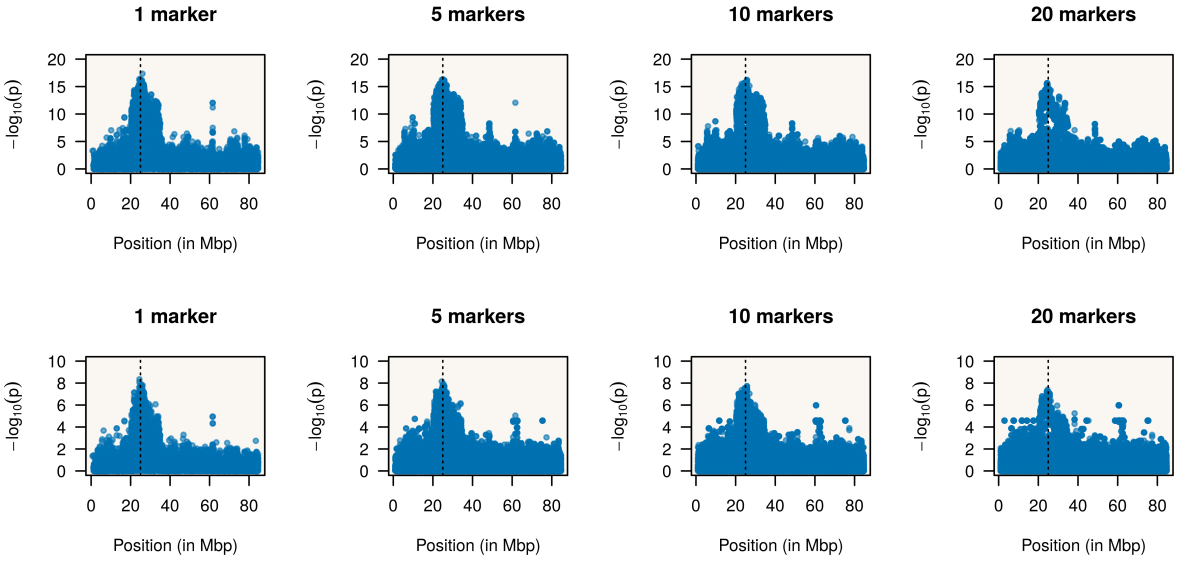

Figure 2 Identification of tag haplotypes for the *PLAG1* locus. The upper and lower panels present results for the fixed and mixed model analyses, respectively. Vertical dashed lines indicate the mid-point position (CHR14:25004732) of candidate causal variants reported by Karim et al. (2011).

Table 7 Candidate quantitative trait nucleotides underlying associations on the *PLAG1* chromosomal domain

Variant name	Position	Functional prediction	Included in the Illumina® BovineHD panel	Reference allele	Candidate *q*	Candidate *Q*
rs110092040	24973953	Downstream variant	No	T	C	T
ss319607399	24974221	Downstream variant	No	A	G	A
ss319607400	24974811	Downstream variant	No	A	G	A
rs109231213	25003338	3'-UTR variant	No	C	G	C
ss319607401	25006125	3'-UTR variant	No	T	C	T
rs109815800	25015640	Intron variant	Yes	G	T	G
ss319607405	25052396	Promoter variant	No	$(CCG)_{11}$	$(CCG)_{9}$	$(CCG)_{11}$
ss319607406	25052440	Promoter variant	No	G	A	G

Table 8. Haplotype diversity at the CHR14:24973324-25015640 *PLAG1* locus in the Bovine HapMap data

Breed	ATACCTT	ATATCTT	ATATTCT	AGATCCT	AGATTCT	AGATTCG	AGGTTCT	GGATTCT	GGGTTCT	GGGTTCG(*Q*)
Blonde d'Aquitaine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Normande	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Norwegian Red	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Red Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0455	0.9545
Holstein	0.0070	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9930
Angus	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0106	0.0000	0.9894
Limousin	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0196	0.0098	0.9706
Charolais	0.0000	0.0000	0.0000	0.0000	0.0250	0.0000	0.0000	0.0000	0.0125	0.9625
Senepol	0.0417	0.0000	0.0000	0.0000	0.0833	0.0000	0.0000	0.0000	0.1250	0.7500
Wagyu	0.0000	0.0000	0.0000	0.0000	0.1538	0.0000	0.0000	0.0000	0.3077	0.5385
Hereford	0.0000	0.0000	0.0000	0.0000	0.0789	0.0000	0.0000	0.0789	0.0526	0.7895
Guernsey	0.0000	0.0000	0.0000	0.0000	0.1667	0.0000	0.0000	0.1190	0.5714	0.1429
Simmental (Fleckvieh)	0.0000	0.0000	0.0000	0.0000	0.3500	0.0000	0.0000	0.1500	0.3000	0.2000
Piedmontese	0.0000	0.0000	0.0000	0.0000	0.1875	0.0000	0.0000	0.3333	0.4792	0.0000
Montbéliard	0.0000	0.0000	0.0000	0.0000	0.4000	0.0000	0.0000	0.2000	0.4000	0.0000
Jersey	0.0000	0.0000	0.0000	0.0000	0.3913	0.0000	0.0000	0.2500	0.2283	0.1304
Romagnola	0.0000	0.0294	0.0000	0.2941	0.4412	0.0000	0.0000	0.0588	0.1618	0.0147
Brown Swiss (Braunvieh)	0.0000	0.0000	0.0000	0.0208	0.2917	0.0000	0.3958	0.1667	0.1250	0.0000
N'Dama	0.0000	0.0000	0.0000	0.0000	0.9167	0.0000	0.0417	0.0208	0.0208	0.0000
Lagunaire	0.0000	0.0000	0.0000	0.0000	0.3000	0.0000	0.0000	0.7000	0.0000	0.0000
Nellore	0.8000	0.1143	0.0143	0.0000	0.0143	0.0000	0.0000	0.0000	0.0143	0.0429
Gyr	0.7167	0.2167	0.0333	0.0000	0.0333	0.0000	0.0000	0.0000	0.0000	0.0000
Beefmaster	0.1042	0.0000	0.0000	0.0000	0.0625	0.0000	0.0000	0.0000	0.0000	0.8333
Santa Gertrudis	0.0429	0.0000	0.0000	0.0000	0.0000	0.0143	0.0000	0.0143	0.0143	0.9143
Brahman	0.2959	0.0306	0.0000	0.0000	0.0204	0.0000	0.0000	0.0000	0.0204	0.6327
Brangus	0.0000	0.0385	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	0.9231
Sheko	0.3889	0.1667	0.0833	0.0000	0.2500	0.0000	0.0833	0.0000	0.0278	0.0000
Buffalo	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gaur	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Yak	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Haplotypes were constructed based on Illumina® BovineHD data of markers BovineHD1400007249 (rs110243083), BovineHD1400007250 (rs136888475), BovineHD1400007253 (rs109636480), BovineHD1400007254 (rs135404594), BovineHD1400007257 (rs134286310), BovineHD1400007258 (rs135538206) and BovineHD1400007259 (rs109815800).

5. Influence of local recombination rates on the estimates of age of selection

Since the age of the *Q* selective sweep was initially dated based on average recombination rates, we re-analyzed the data using sex-specific high resolution recombination maps reported by Ma et al. (2015) (available at: http://datadryad.org/resource/doi:10.5061/dryad.q2q84). These maps were generated from a large pedigree of 186,927 three-generation families and built over ~8.5 million observed recombination events. Genetic distances were first computed using Haldene's map function (HALDENE, 1919), then missing values were linearly interpolated. As shown in Table 9, the use of local or average recombination rates provided very similar results.

Table 9. Estimates of 95% confidence intervals for the age of the *Q* selective sweep using different genetic maps

Population	Average (1.23 Mbp ~ 1 cM)	Female[a]	Male[a]
European *B. taurus*	305 - 475	250 - 410	250 - 405
Japanese *B. taurus*	90 - 195	100 - 205	100 - 205
Brahman	75 - 170	70 - 165	70 - 165
Nellore	30 - 100	30 - 95	30 - 100

[a]Sex-specific genetic maps were generated from recombination rates reported by Ma et al. (2015).

6. References

ALLAIS-BONNET, A.; GROHS, C.; MEDUGORAC, I.; KREBS, S.; DJARI, A.; GRAF, A.; FRITZ, S.; SEICHTER, D.; BAUR, A.; RUSS, I.; BOUET, S.; ROTHAMMER, S.; WAHLBERG, P.; ESQUERRÉ, D.; HOZE, C.; BOUSSAHA, M.; WEISS, B.; THÉPOT, D.; FOUILLOUX, M. N.; ROSSIGNOL, M. N.; VAN MARLE-KÖSTER, E.; HREIÐARSDÓTTIR, G. E.; BARBEY, S.; DOZIAS, D.; COBO, E.; REVERSÉ, P.; CATROS, O.; MARCHAND, J. L.; SOULAS, P.; ROY, P.; MARQUANT-LEGUIENNE, B.; LE BOURHIS, D.; CLÉMENT, L.; SALAS-CORTES, L.; VENOT, E.; PANNETIER, M.; PHOCAS, F.; KLOPP, C.; ROCHA, D.; FOUCHET, M.; JOURNAUX, L.; BERNARD-CAPEL, C.; PONSART, C.; EGGEN, A.; BLUM, H.; GALLARD, Y.; BOICHARD, D.; PAILHOUX, E.; CAPITAN, A. Novel insights into the bovine polled phenotype and horn ontogenesis in Bovidae. **PLOS ONE**, v. 8, e63512, 2013.

BOITARD, S.; BOUSSAHA, M.; CAPITAN, A.; ROCHA, D.; SERVIN, B. Uncovering adaptation from sequence data: Lessons from genome resequencing of four cattle breeds. **Genetics**, v. 203, p. 433-450, 2016.

BROWNING, B. L.; BROWNING, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **American Journal of Human Genetics**, v. 84, p. 210-223, 2008.

CARVALHEIRO, R.; BOISON, S. A.; NEVES, H. H. R.; SARGOLZAEI, M.; SCHENKEL, F. S.; UTSUNOMIYA, Y. T.; PÉREZ-O'BRIEN, A. M.; SÖLKNER, J.; MCEWAN, J. C.; VAN TASSELL, C. P.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution (GSE)**, v. 46, 69, 2014.

HALDANE, J. B. S. The combination of linkage values and the calculation of distances between the loci of linked factors. **Journal of Genetics**, v. 8, p. 299-309, 1919.

KARIM, L.; TAKEDA, H.; LIN, L.; DRUET, T.; ARIAS, J. A. C.; 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, p. 405-413, 2011.

MA, L.; O'CONNELL, J. R.; VANRADEN, P. M.; SHEN, B.; PADHI, A.; SUN, C.; BICKHART, D. M.; COLE, J. B.; NULL, D. J.; LIU, G. E.; DA, Y.; WIGGANS, G. R. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. **PLOS Genetics**, v. 11, e1005387, 2015.

MEDUGORAC, I.; SEICHTER, D.; GRAF, A.; RUSS, I.; BLUM, H.; GÖPEL, K. H.; ROTHAMMER, S.; FÖRSTER, M.; KREBS, S. Bovine polledness - an autosomal dominant trait with allelic heterogeneity. **PLOS ONE**, v. 7, e39477, 2012.

PICCOLI, M. L.; BRACCINI, J.; CARDOSO, F. F.; SARGOLZAEI, M.; LARMER, S. G.; SCHENKEL, F. S. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. **BMC Genetics**, v.15, 157, 2014.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v. 15, 478, 2014.

UTSUNOMIYA, A. T. H.; SANTOS, D. J. A.; BOISON, S. A.; UTSUNOMIYA, Y. T.; MILANESI, M.; BICKHART, D. M.; AJMONE-MARSAN, P.; SÖLKNER, J.; GARCIA, J. F.; DA FONSECA, R.; DA SILVA, M. V. G. B. Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. **BMC Genomics**, v. 17, 705, 2016.

VENTURA, R. V.; LU, D.; SCHENKEL, F. S.; WANG, Z. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. **Journal of Animal Science**, v. 92, p. 1433-1444, 2014.

YANG, J.; ZAITLEN, N. A.; GODDARD, M. E.; VISSCHER, P. M.; PRICE, A. L. Advantages and pitfalls in the application of mixed-model association methods. **Nature Genetics**, v. 46, p. 100-106, 2014.