THE UNIVERSITY
of EDINBURGH

Genetics of disease resistance: application to bovine Tuberculosis

Smaragda Tsairidou

Doctor of Vet. Med. (Thessaloniki)

M.Sc (Edinburgh)

This thesis is presented for the degree of
Doctor of Philosophy

College of Medicine and Veterinary Medicine
The Roslin Institute and Royal (Dick) School of Veterinary Studies
University of Edinburgh

2015

This Thesis is dedicated to my supervisor,

Professor Stephen C. Bishop (1960-2015)

Declaration

I hereby declare that I am the author of this thesis and that I did all the work described herein, unless otherwise specified. This work has not been submitted for any other degree or professional qualification except as specified. I have read and understood The University of Edinburgh guidelines on Plagiarism and declare that this written dissertation is all my own work except where I indicate otherwise by proper use of quotes and references. This thesis is an account of work conducted by me whilst studying for the degree of Doctor of Philosophy at the University of Edinburgh.

Smaragda Tsairidou

Date: 30.9.15

Acknowledgements

I am deeply grateful to my supervisor, Professor Stephen Bishop, for his invaluable help and advice. Steve was an inspiring mentor and his teaching, guidance and encouragement have been of immense value to me and my future career. He is greatly missed.

I am deeply grateful to my supervisor, Professor John Woolliams, for supervising and supporting this project. John, I would like to express my sincere thanks and appreciation for your precious support and advice. Your supervision and guidance have been of crucial importance for the completion of my PhD and have helped me further develop my skills as a researcher.

My appreciation also goes to my supervisor Professor Georgios Banos, for the useful discussions and his helpful advice and support through the entire course of my studies. I am thankful to Dr Ricardo Pong-Wong for his valuable contribution to my project. Ricardo, I would call you my 4th supervisor, thank you for your help and for a great friendship.

I would like to thank the rest of my Thesis committee, Jo Stevens and Pam Wiener, my internal examiner Dr Ian Handel, and my external examiner Professor Erling Strandberg, for their assistance and their insightful comments. I would also like to thank our project collaborators from the Agri-Food and Biosciences Institute, in Northern Ireland, and in particular Dr Adrian Allen and Dr Robin Skuce, and I would also like to thank Professor Liz Glass from the Roslin Institute and Professor Mike Coffey from SRUC. I acknowledge the Department of Agriculture and Rural Development for access to Animal and Public Health Information System data,

Science Foundation Ireland 09/IN.1/B2642, access to the DAFF dataset, and the Department for Environment Food & Rural Affairs for access to bTB skin test data.

This work was conducted under the Principal‘s Career development PhD Scholarship, at the College of Medicine and Veterinary Medicine, University of Edinburgh. I am grateful for being awarded Associate Fellowship of The Higher Education Academy and for receiving the Edinburgh Teaching Award. I would like to thank the director of postgraduate studies at the Roslin Institute, Professor Bernadette Dutia, the director of veterinary teaching at the Royal (Dick) School of Veterinary Studies, Professor Susan Rhind, the Institute for Academic Development and in particular my mentor Dr Miesbeth Knottenbelt. I am thankful for the Greek State Scholarship Foundation award and for their financial support. I am also thankful to the British Society of Animal Science (BSAS) for the Alan Robertson award, the Roslin Institute and the Royal (Dick) School of Veterinary Studies for the Birrell-Gray Travelling Scholarship, and the Genetic Society for the Junior Scientist Conference Grand.

Last but not least, thank you to my dear friends and colleagues at the Division of Genetics and Genomics at the Roslin Institute. Especially I would like to thank Oswald Matika, Enrique Sanchez Molano, Chris Pooley, and Fiona Milton for their help and friendship. Thank you to David Telford for his valuable encouragement and understanding. I am deeply grateful to my beloved family, my Parents Anastasia and Ilias and my uncle George.

Thank you to all who contributed directly or indirectly to this project.

List of Publications

Research articles (peer-reviewed)

S. Tsairidou, J. A. Woolliams, A. R. Allen, R. A. Skuce, S. H. McBride, D. M. Wright, M. L. Bermingham, R. Pong-Wong, O. Matika, S. W. J. McDowell, E. J. Glass, S. C. Bishop. (2014) Genomic Prediction for Tuberculosis Resistance in Dairy Cattle. Plos One, Vol. 9, Issue 5, e96728 (Based on Chapter 2).

S. Tsairidou, J.A. Woolliams, A.R. Allen, R.A. Skuce, S.H. McBride, R. Pong-Wong, O. Matika, E.K. Finlay, D.P. Berry, D.G. Bradley, S.W.J. McDowell, E.J. Glass, S.C. Bishop. (2016) A meta-analysis for bovine Tuberculosis resistance in dairy cattle. Manuscript under preparation (Based on Chapters 4 and 5).

S. Tsairidou, S. Brotherstone, M. P. Coffey, S.C. Bishop, J.A. Woolliams. (2016) Quantitative Genetic Analysis of the bTB Diagnostic Single Intradermal Comparative Cervical Test (SICCT). Manuscript submitted (Based on Chapter 6).

Conference Proceedings (peer-reviewed)

S. Tsairidou, M. L. Bermingham, S. C. Bishop, J. A. Woolliams, A. R. Allen, S. H. McBride, D. M. Wright, R. A. Skuce, R. Pong-Wong, O. Matika, S. W. J. McDowell, E. J. Glass. (2013) Genomic prediction for tuberculosis resistance in dairy cattle. Proceedings of the British Society of Animal Science, BSAS Annual Meeting 2013, Nottingham, UK.

S. Tsairidou, J.A. Woolliams, A.R. Allen, R.A. Skuce, S.H. McBride, R. Pong-Wong, O. Matika, E.K. Finlay, D.P. Berry, D.G. Bradley, S.W.J. McDowell, E.J. Glass, S.C. Bishop. (2014) A meta-analysis for bovine tuberculosis resistance in dairy cattle. Proceedings, 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada.

S. Tsairidou, J.A.Woolliams, A.R. Allen, R.A. Skuce, S.H. McBride, D.M. Wright, M.L. Bermingham, R. Pong-Wong, O. Matika, C. M. Pooley, S.W.J. McDowell, E.J. Glass, S.C. Bishop. (2014) A Heterozygote Advantage Analysis For Tuberculosis Resistance In Dairy Cattle. VI International *M. bovis* Conference, Cardiff, Wales.

S. Tsairidou, J.A. Woolliams, A.R. Allen, R.A. Skuce, S.H. McBride, R. Pong-Wong, O. Matika, E.K. Finlay, D.P. Berry, D.G. Bradley, S.W.J. McDowell, E.J. Glass, S.C. Bishop. (2015) A meta-analysis for bovine tuberculosis resistance in dairy cattle. Proceedings of the British Society of Animal Science, BSAS Annual Meeting 2015, Chester, UK.

Popular Science

VIth International *Mycobacterium bovis* Conference. (2015) Genetics Society News, Issue 72.

Table of Contents

Abstract

Bovine Tuberculosis (bTB) is a disease of significant economic importance, being one of the most persistent animal health problems in the UK and the Republic of Ireland and increasingly constituting a public health concern especially for the developing world. Limitations of the currently available diagnostic and control methods, along with our incomplete understanding of bTB transmission, prevent successful eradication. This Thesis addresses the development of a complementary control strategy which will be based on animal genetics and will allow us to identify animals genetically predisposed to be more resistant to disease. Specifically, the aim of my PhD project is to investigate the genetic architecture of resistance to bTB and demonstrate the feasibility of whole genome prediction for the control of bTB in cattle. Genomic selection for disease resistance in livestock populations will assist with the reduction of the in herd-level incidence and the severity of potential outbreaks.

The first objective was to explore the estimation of breeding values for bTB resistance in UK dairy cattle, and test these genomic predictions for situations when disease phenotypes are not available on selection candidates. Through using dense SNP chip data the results of Chapter 2 demonstrate that genomic selection for bTB resistance is feasible ($h^2 = 0.23(SE = 0.06)$) and bTB resistance can be predicted using genetic markers with an estimate of prediction accuracy of $r(g, \hat{g}) = 0.33$ in this data. It was shown that genotypes help to predict disease state ($AUC \approx 0.58$) and animals lacking bTB phenotypes can be selected based on their genotypes. In Chapter 3, a novel approach is presented to identify loci displaying heterozygote

(dis)advantage associated with resistance to *M. bovis*, hypothesising underlying non-additive genetic variation, and these results are compared with those obtained from standard genome scans. A marker was identified suggesting an association between locus heterozygosity and increased susceptibility to bTB i.e. a heterozygote disadvantage, with the heterozygotes being significantly more in the cases than in the controls ($x^2 = 11.50, p<0.001$).

Secondly, this thesis focused on conducting a meta-analysis on two dairy cattle populations with bTB phenotypes and SNP chip genotypes, identifying genomic regions underlying bTB resistance and testing genomic predictions by means of cross-validation. In Chapter 4, exploration of the genetic architecture of the trait revealed that bTB resistance is a moderately polygenic, complex trait with clusters of causal variants spread across a few major chromosomes collectively controlling the trait. A region was identified on chromosome 6, putatively associated with bTB resistance and this chromosome as a whole was shown to contribute a major proportion ($h_c^2= 0.051$) of the observed variation in this dataset. Genomic prediction for bTB was shown to be feasible even when only distantly related populations are combined ($r(g,\hat{g})=0.33$ $(SE = 0.05)$), with the chromosomal heritability results suggesting that the accuracy arises from the SNPs capturing linkage disequilibrium between markers and QTL, as well as additive relationships between animals (*~80%* of estimated genomic h^2 is due to relatedness). To extend the analysis, in Chapter 5, high density genotypes were inferred by means of genotype imputation, anticipating that these analyses will allow the identification of genomic regions associated with bTB resistance more closely, and that would increase the prediction accuracy. Genotype imputation was successful, however,

using all imputed genotypes added little information. The limiting factor was found to be the number of animals and the trait definitions rather than the density of genotypes.

Thirdly, a quantitative genetic analysis of actual Single Intradermal Comparative Cervical Test (SICCT) values collected during bTB herd testing was conducted aiming to investigate if selection for bTB resistance is likely to have an impact on the SICCT diagnostic test. This analysis demonstrated that the SICCT has a negligibly low heritability (h^2*=0.0104 (SE = 0.0032)*) and any effect on the responsiveness to the test is likely to be small.

In conclusion, breeding for disease resistance in livestock is feasible and we can predict the risk of bTB in cattle using genomic information. Further, putative QTLs associated with bTB resistance were identified, and exploration of the genetic architecture of bTB resistance revealed a moderately polygenic trait. These results suggest that given that larger datasets with more phenotyped and genotyped animals will be available, we can breed for bTB resistance and implement the genomic selection technology in breeding programmes aiming to improve the disease status and overall health of the livestock population. Using the genomics this can be continued as the epidemic declines.

Lay summary

Bovine Tuberculosis (bTB) is one of the most persistent animal health problems in many countries around the world and remains a major challenge for the UK and the Republic of Ireland (RoI) despite the on-going eradication programmes. Limitations in the currently available diagnostic and control methods hinder eradication and therefore, it is becoming increasingly clear that complementary strategies will be needed to control bTB. My PhD project addresses the hypothesis that genetic selection for disease resistance in the light of genomic advances, may offer a complementary strategy for the control of bTB in cattle, by reducing infection risks. Genomic selection is a new technology that allows to identify animals that are genetically predisposed to be more resistant to disease by utilising information of genetic markers (Single Nucleotide Polymorphisms, SNPs) spread across the genome, without requiring regular collection of phenotypic information or knowledge of the exact genes controlling the trait. Such an approach will contribute to a reduction in herd-level incidence as well as a reduction in the severity of an outbreak.

More specifically, the aims of my PhD project were to study the genetic architecture of resistance to bTB, and to explore the feasibility of genomic selection for livestock populations that will be more resistant to disease. My results have demonstrated that genomic selection for bTB resistance is feasible, and using genetic markers, bTB resistance can be predicted for situations when disease phenotypes are not available. Furthermore, a single-SNP approach capturing non-additive genetic variation identified a locus suggesting an association between locus heterozygosity

and increased susceptibility, however, it was found that it is unlikely that bTB is controlled by a single gene. Therefore, to extend my analysis, two populations only distantly related were combined in a meta-analysis. Genomic prediction for bTB was shown to be feasible even when only distantly related populations are combined and a genomic region underlying bTB resistance was identified. Exploration of the genetic architecture of the trait revealed that bTB resistance is a moderately polygenic, complex trait with clusters of causal variants spread across a few major chromosomes collectively controlling the trait. High density genotypes inferred using cost effective methods such as genotype imputation will allow for the SNPs to be more closely linked to the QTLs (Quantitative Traits Loci). However, in the present data the limiting factor was found to be the number of animals and the trait definitions rather than the density of genotypes. Lastly, a quantitative genetic analysis of actual Single Intradermal Comparative Cervical Test (SICCT) values, which is the test used for bTB diagnosis in the UK, was conducted. This analysis demonstrated that the SICCT has a negligibly low heritability and any effect of the selection for bTB resistance, on the responsiveness to the diagnostic test, is likely to be small.

In conclusion, breeding for disease resistance in livestock is feasible and we can predict the risk of bTB in cattle using genomic information. BTB resistance was shown to be a moderately polygenic trait, however, the prediction accuracy was found to arise from true linkage between the markers and the QTLs, as well as familial relationships between animals. These results suggest that given that larger datasets with more phenotyped and genotyped animals will be available, we can breed for bTB resistance and implement the genomic selection technology in

breeding programmes aiming to improve the disease status and overall health of the livestock population. Using the genomics this can be continued as the epidemic declines.

Chapter 1

General Introduction

1.1 Bovine Tuberculosis (bTB)

Bovine Tuberculosis (bTB) is a bacterial disease caused by *Mycobacterium bovis* (*M. bovis*), an aerobic Gram^{+} bacillus and member of the *M. tuberculosis* complex. Cattle (*Bos taurus*) predominantly become infected through the respiratory route and the main lesions observed are tubercles formed in the lungs and draining lymph nodes (Divers and Peek 2008). BTB is a zoonotic disease and has an impact on animal performance and welfare, causing significant financial losses to the dairy cattle industry worldwide due to production losses and the cost of eradication programmes (Allen et al. 2010).

1.1.1 Overview

BTB is one of the most persistent animal health problems in many countries including the UK and the Republic of Ireland (RoI) (Allen et al. 2010). It is estimated that the total cost of eradication programmes for the years 2010-2011 has exceeded *£275* million (UK and RoI) (Abernethy et al. 2013), and overall, the number of new outbreaks in Great Britain has increased since 1996 (*3.9%* incidence rate, 2015), with no clear evidence of a decline despite the control measures (Fig. 1) (https://www.gov.uk/government/statistics/historic-statistics-notices-on-the-incidence-of-tuberculosis-tb-in-cattle-in-great-britain-2015). Whilst human tuberculosis is usually caused by *Mycobacterium tuberculosis*, humans can also get

infected by *M. bovis*, and therefore bTB is a zoonotic disease and constitutes a growing public health concern worldwide. Zoonotic transmission occurs through unpasteurised milk consumption (Buddle et al. 2003) and for the developing world bTB has been listed as the fourth most important livestock disease (Perry et al. 2002) with *M. bovis* causing *10-15%* of TB in humans (Michel et al. 2010).

Mycobacteria of the tuberculosis complex are among the most persistent pathogens known, having developed impressive mechanisms for evolution and adaptation (Raman et al. 2009). Garnier et al. (2003) reported that *M. bovis* has a sequence length of 4,345,492 bp with *3,952* protein encoding genes. *M. tuberculosis*, the causative agent of TB in humans, and *M. bovis* are distinct lineages of the *Mycobacterium tuberculosis* complex, and *M. bovis* has most likely descended from *M. tuberculosis* after a series of deletion mutations in its genome, although the average sequence divergence between them is low (*<0.05%*) (Brosch et al. 2001; Garnier et al. 2003; Gonda et al. 2006; Hewinson et al. 2006). *M. bovis* shows reduced diversity in the UK and the RoI, where the dominant clonal complex is the Eu1 carrying the SB0140 spoligotype (Smith et al. 2006; Smith et al. 2011). Spoligotyping is a molecular technique for genotyping and classifying *M. tuberculosis* complex strains through identifying polymorphisms in the direct repeat region of their genome (Smith et al. 2006). A possible explanation suggested for the extreme clonality observed, is the population bottleneck introduced by the bTB eradication programmes (Smith et al. 2006).

1.1.2 On bTB transmission and pathogenesis

BTB is a chronic, primarily respiratory infectious disease. Cow to cow transmission mainly occurs through respiratory secretions. *M. bovis* is an intracellular pathogen and after being inhaled, aerosol droplets carrying Mycobacteria reach the lungs where the primary target-cells are the alveolar macrophages (Fig. 2) (Koul et al. 2004). There, Mycobacteria act as a type of intracellular parasite managing to survive within the macrophages by disrupting cell-signalling pathways, overriding immune response (Koul et al. 2004) and establishing a chronic infection. The typical lesions observed are tubercles formed mainly in the lungs and the lymph nodes that drain the region (Divers and Peek 2008). Tubercles contain Mycobacteria and immune cells enclosed within multiple layers of fibrous tissue, and their structure assists in preventing spread of the infection (Fayyazi et al. 2000; Saunders et al. 2000; Cassidy et al. 2005). Although inhalation is the main route of transmission, younger cattle may also be infected by ingestion, through consumption of infected milk (Divers and Peek 2008).

Transmission of bTB can also occur from a wildlife reservoir into cattle populations, which sustains infection. Therefore, the wildlife reservoir is one of the major factors that hinder eradication. The Eurasian badger (*Meles meles)* is the main wildlife reservoir for the disease in the UK and Ireland (https://www.gov.uk/government/publications/2010-to-2015-government-policy-bovine-tuberculosis-bovine-tb). Other wildlife species such as the red deer (*Cervus elaphus*) and the possums (*Trichosurus vulpecula*) (New Zealand) are wildlife reservoirs of the disease around the world. Furthermore, pets, and specifically cats

and dogs, can become infected with *M. bovis* mainly through drinking unpasteurised infected milk or contact with infected animals (farm animals or wildlife) (https://www.gov.uk/government/publications/bovine-tuberculosis-tb-in-domestic-pets), and there are reports on transmission to humans through contact with cats infected with *M. bovis* (de Lisle et al. 1990; Monies et al. 2006). Lastly, multidrug resistant *M. bovis* strains and the possibility of human-to-human transmission introduce new challenges to bTB control (Cosivi et al. 1998). *M. bovis* has a capacity for causing major epidemics as it is indicated by its basic reproductive number being greater than one (Cox et al. 2005), and therefore it is necessary to develop and employ efficient diagnostic and control strategies.

1.1.3 Diagnosis of bTB

The main clinical manifestations of bTB are emaciation, chronic cough, lymph node enlargement, and udder infection (Divers and Peek 2008). However, bTB is a chronic disease and the clinical symptoms can be very rarely observed in veterinary practice. Thus, there is a range of approaches that are available and used for diagnosing bTB including the tuberculin skin test, post-mortem examination, bacteriology and histological analysis, and interferon-gamma testing. Important epidemiological parameters that indicate the qualities of a diagnostic test are its sensitivity and specificity, representing the power of the test to correctly identify the infected and the healthy individuals as such, respectively. Limitations in the sensitivity and specificity of the available diagnostic methods along with our incomplete understanding of bTB transmission hinder successful eradication. In the UK, the cattle industry has traditionally relied on diagnosis through compulsory

tuberculin skin testing and abattoir carcass inspections (de la Rua-Domenech et al. 2006).

Skin testing using the Single Intradermal Comparative Cervical Test (SICCT) has been the cornerstone of bTB surveillance: animals with a positive outcome to the test are culled and movement restrictions are applied to the herd (Allen et al. 2010). The test is based on the interpretation of the local inflammatory response and skin swelling due to the delayed-type hypersensitivity reaction to the tuberculin antigen that is inoculated (i.e. Purified Protein Derivative, PPD) (Allen et al. 2010). Two important elements of the test are the test design and the re-testing intervals. Historically, several variations of the test have been developed, including the single and the double dose tests, intradermal or subcutaneous administration, cervical inoculation or at the caudal fold, use of the mammalian or the bovine PPD (VLA Weybridge or Prionics ID-Lelystad), a single tuberculin or a comparative test. Lesslie et al. (1975) demonstrated through their classical tuberculin testing trials that the bovine tuberculin, and specifically the comparative version of the test, improved the specificity of the test and its power to discriminate between tuberculous and non-tuberculus animals, in herds where *M. avium*, skin tuberculosis and Johne's disease were also present, with a number of animals in the trial being vaccinated for Johne's disease. The comparative version of the test has been employed in the UK in order to take into account the non-specific reactions to *M. avium sbsp avium,* which is also present in the environment and might cause false positives. Over the years, a complex re-testing protocol following identification of reactors or inconclusive reactors in a herd has been developed. The basis of this protocol is the *60*-days interval between subsequent tests, in order to avoid the suppressive effect of the test

itself causing reduced reactivity to the tuberculin for a subsequent time interval (i.e. the "desensitisation effect") (Radunz and Lepper 1985; Kerr et al. 1946; Monaghan et al. 1994; Doherty et al. 1995; Thom et al. 2004). However, the underlying mechanisms of desensitisation, and when reactivity returns to its previous levels, along with how these could be exploited in the modern re-testing protocol to shorten those intervals for a more stringent test, are poorly studied. Obligatory testing intervals range from *6* months (Intensive Action Areas) to *4* years, depending on the county disease risk (from 1.1.2013). However, while SICCT as applied has very good specificity (*>99% Sp*), it suffers from insufficient sensitivity (*~55-70% Se*) (i.e. allowing for false negatives) (Neill et al. 1994; Olea-Popelka et al. 2004; De la Rua-Domenech et al. 2006). Although SICCT has been effective at the whole-herd level, i.e. in diagnosing infected herds, it has not been very successful in diagnosing the infection status at the level of the individual animal, and cattle that were negative to the test although infected (i.e. false negatives), maintain an infectious challenge in the herd.

At post-mortem examination in the abattoir the main lesions observed are tubercles that can be found in the lungs and the lymph nodes. However, carcass inspection also has insufficient sensitivity (*~30-50% Se*) (Neill et al. 1994; Olea-Popelka et al. 2004; De la Rua-Domenech et al. 2006). Laboratory confirmation of infection for the tuberculin test reactors or suspect abattoir lesions is based on a combination of bacteriology, histological analysis and culture of tissue samples. However, this is problematic due to the highly specific requirements of the slow-growing Mycobacteria *in vitro* (Cosivi et al. 1998).

One of the most recent diagnostic tools is the interferon-gamma (IFN-γ) test, which is an enzyme linked immune-sorbent assay (ELISA) that detects IFN-γ in whole blood, released as a reaction to tuberculin (Allen et al. 2010). Although the IFN-γ test has reportedly higher sensitivity than the standard interpretation SICCT, it has substantially lower specificity (*~96% Sp*) (Downs et al. 2011). Therefore, its use has been suggested in combination with the skin test, as a complementary diagnostic test (Sheridan 2011).

1.1.4 bTB control strategies

Control of bTB remains a major challenge despite the on-going eradication programmes in the UK. Eradication strategies are compromised by the presence of the wildlife reservoir. Studies on the effectiveness of culling badgers in the UK to reduce bTB prevalence in cattle (the Randomised Badger Culling Trial (RBCT), 1998-2005) have shown both positive and negative effects. Although beneficial effects were observed within the badger culling areas, there were negative effects on the adjoining lands, with an increase in cattle bTB incidence possibly due to the social perturbation and immigration of badgers caused by the culling (Cox et al. 2005; Donnelly et al. 2007; Jenkins et al. 2010; Sheridan 2011). Moreover, recent studies have argued that control strategies that focus on a single route of transmission are not likely to be very successful (Brooks-Pollock et al. 2014).

One strategy that could complement eradication is vaccination. Bacillus Calmette Guerin (BCG) vaccine, initially developed for humans, is based on an attenuated *M. bovis* strain containing numerous deletions that make it non-virulent (Buddle et al. 2003; Behr et al. 2015). However, BCG vaccination is precluded

because it is not fully protective, while vaccinated animals are currently indistinguishable from naturally infected animals using the standard tuberculin tests (Cosivi et al. 1998; Buddle et al. 2003; Hewinson et al. 2006). The possibility of developing tests that will allow Differentiating Infected from Vaccinated Animals (DIVA) is under assessment and currently, there are no licenced cattle vaccines for use against bTB in the UK.

Identification of infected animals using the available diagnostic methods is further complicated by the various demonstrations of bTB infection associated with the progress of disease, that have an impact on the responsiveness to the diagnostic skin test and the effectiveness of carcass inspections. For example, although some animals react to the skin test, they do not show any detectable lesions in abattoir inspections (Non-Visible Lesions reactors, NVLs) (Radunz et al. 1985; Comer et al. 1994; Doherty & Cassidy 2002). This is an intrinsic problem of the abattoir inspection-based diagnosis, as identification of a bTB case cannot solely rely on this method. Moreover, there are Non-Specific Reactors (NSRs) that complicate reaction to the SICCT and the interpretation of test results, which can be due to a number of factors including natural, experimental, or vaccinal exposure to paratuberculosis (see below), *M. avium*, skin TB (*M. microti*), or other environmental mycobacteria that share proteins with the bovine tuberculin (Monaghan et al. 1994). Further, there is a time period of *30-50* days required post infection for the animal to develop reactivity to SICCT (Monaghan et al. 1994) and *21-35* days to develop detectable reaction to the IFN-γ test (Dean et al. 2005). Thus, animals that are recently infected and are still at initial stages of infection fail to respond to the SICCT although infected (Vallee and Panisset 1920; Francis 1947). Anergy describes animals that although they have

severe and generalised disease, they do not react to tuberculin due to various reasons including parturition (due to the periparturient immunosuppression), stress, treatment with glucocorticoids, BVD virus infection, and malnutrition (Kerr 1949; Buddle et al. 1994; Skuce et al. 2011). Finally, latency, which has mostly been studied in humans, is a condition where the mycobacteria remain inactive for a long period of time. Individuals with latent infection show no clinical signs and they are not able to spread the disease unless the disease becomes active (Flynn et al. 2001).

The outcome of the diagnostic tests can be compromised by the presence of simultaneous infections. Paratuberculosis (Johne's disease), caused by *Mycobacterium avium sbsp paratuberculocis* is also prevalent in the UK (*~34.7%* prevalence; http://archive.defra.gov.uk/), and cattle infected with Paratuberculosis can be responsive to the bovine and avian tuberculins (Lesslie et al. 1975; Monaghan et al. 1994). Vaccination against it might be interfering with the interpretation of the tuberculin skin test (Köhler et al. 2001), while paratuberculosis infection also decreases the specificity and sensitivity of the IFN-γ test (Alvarez et al. 2009). Moreover, recent studies have shown that when coinfection with parasites occurs, and in particular with the liver fluke (*Fasciola hepatica*), the bTB skin test might be less sensitive (Claridge et al. 2012).

BTB diagnosis and control is an ongoing challenge for the UK cattle industry. Therefore, in a situation where conventional control strategies have not been effective, it is becoming increasingly clear that complementary strategies will be needed to control bTB.

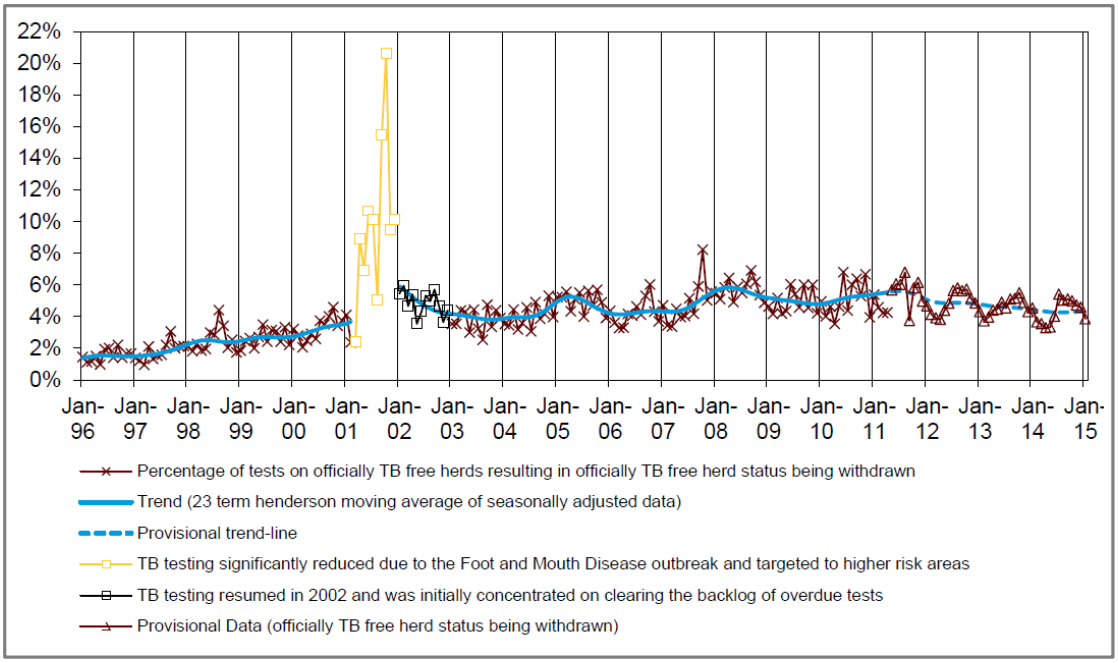


Figure 1. The change of bTB incidence in GB between the years 1996 to 2015. In the graph we see the number of new incidents of bTB leading to the withdrawal of Officially bTB Free (OTF) herd status, as a percentage of tests carried out in OTF herds. The spike observed in 2001 is due to the suspension of bTB testing during the Foot and Mouth Disease (FMD) epidemic. Overall, the incidence rate was at its peak in 2008. (https://www.gov.uk/government/statistics/historic-statistics-notices-on-the-incidence-of-tuberculosis-tb-in-cattle-in-great-britain-2015).

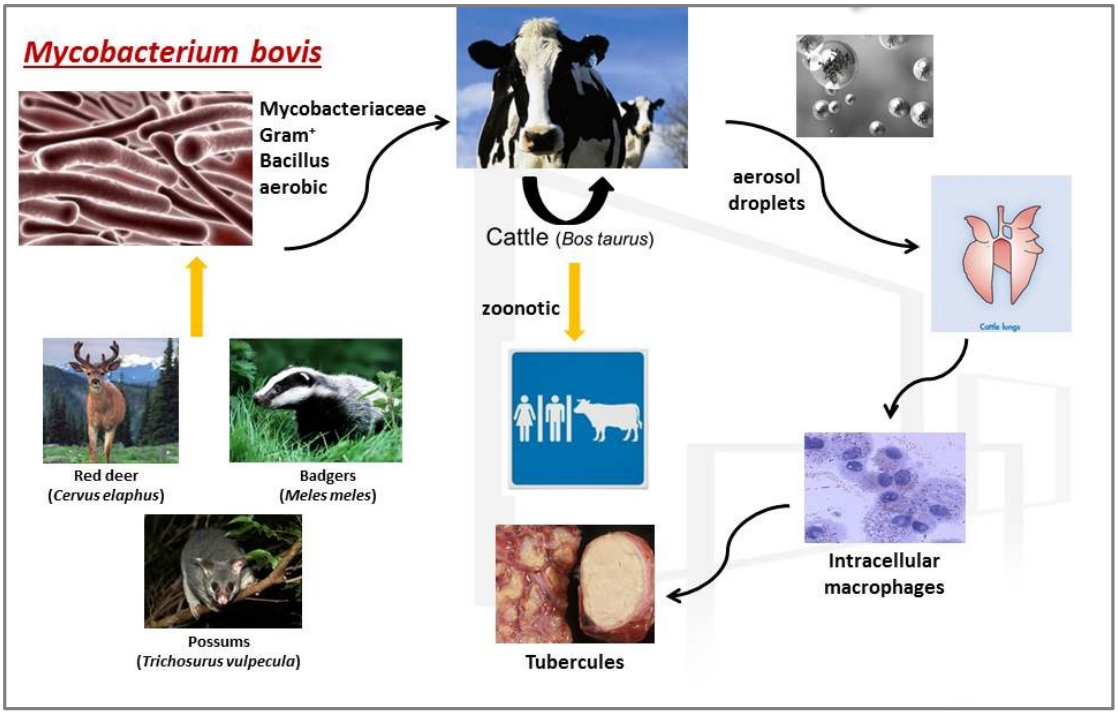


Figure 2: Transmission of *M. bovis*.

1.2 Genetic selection for disease resistance

1.2.1 Genomic prediction - overview

Following exposure to *M. bovis* only a proportion of animals develop disease, implying variability among individuals in terms of their response to infection (Pollock et al. 2002). Various studies have confirmed that the variation in bTB resistance among dairy cattle is in part genetic and it is exploitable, and have demonstrated moderate to strong heritability (Bermingham et al. 2009; Brotherstone et al. 2010; Tsairidou et al. 2014) (Table 1). Moreover, variation has been found among cattle family lines (Phillips et al. 2002), and between Holstein cattle, zebus (*Bos indicus*) and zebus x Holstein crosses, with Holsteins being more susceptible than zebus or crosses (Ameni et al. 2007). Thus, a proportion of the observed phenotypic variation can be attributed to host genetics.

Improvement of livestock through selective breeding of the best animals has been long practised by farm animal breeders. Genetic selection utilises both phenotypes (i.e. bTB state) and pedigree information (Henderson 1975; Goddard and Hayes, 2009). Estimated Breeding Values (EBVs) can then be calculated using statistical techniques such as Best Linear Unbiased Prediction (BLUP). However, it requires information on phenotypes or indicator traits which is not always feasible and particularly for traits difficult to measure or when the diagnostic test has imperfect specificity and sensitivity. For example in the case of bTB, phenotypes are difficult to collect as infection can only be confirmed post mortem and the SICCT has imperfect sensitivity. Furthermore, and particularly when within-herd prevalence is low (i.e. $p \sim 6\text{-}10\%$ for bTB), selecting phenotypes is inefficient as exposure to the

pathogen is required to express the resistant genotype. With incomplete exposure to infection some animals do not have the opportunity to express their resistant phenotype (Bishop and Woolliams 2010). Thus, such selection would work only on the subset of animals in herds affected by bTB, or their close relatives, and it would require that the population is undergoing an epidemic. Even then, selection intensity would be low if only a small proportion of herds were affected (Bishop and Woolliams 2010). Lastly, even if selection based on phenotypes were to be undertaken and had some success in controlling the epidemic, this success would reduce the potential for making further progress through a reduced number of informative phenotypes. Therefore, in the case of bTB resistance, it is appealing to be able to identify relatively resistant animals in the absence of phenotypic data from an epidemic.

The idea of Genomic Selection (GS) as established by Meuwissen et al. (2001), introduced a new potential to the world of research towards genetic improvement of livestock (Goddard et al. 2007). In contrast to selection based on routine phenotyping, genomic selection is a technology that addresses the problem of identifying relatively resistant individuals by obtaining EBVs for animals without observing phenotypes. Genomic selection utilises genomic EBVs estimated directly from SNP data rather than pedigree data, calculated as the sum of the effects of genetic markers (Single Nucleotide Polymorphisms, SNPs) across the genome (Hayes et al. 2009; Jia et al. 2012). Regions containing genes controlling quantitative traits, such as disease resistance, are called Quantitative Trait Loci (QTL). The central idea is that markers across the genome will be sufficiently close to the QTLs and will assist predicting the effect of the QTLs through Linkage Disequilibrium

(LD) (Goddard 2009). The most commonly used method for calculating EBVs is Genomic BLUP (GBLUP) (Meuwissen et al. 2001; Daetwyler et al. 2010). The process can be summarised in two steps: Initially, GEBVs are calculated on a training population with both phenotypic and genotypic information e.g. from an epidemic. Then, for the selection candidates, GEBVs can be predicted with a certain accuracy without phenotypic records (Meuwissen et al. 2001), by combining knowledge on their genotypes and the marker effects which have already been calculated in the training population (Hayes et al. 2009; Luan et al. 2009). In other words, through the genomic prediction methodology, EBVs can be estimated by combining knowledge on genotypes of the selection candidates and marker effects, and these can then be used as predictors of disease susceptibility for every animal. The estimation of marker effects relies on the LD between the markers and the QTLs, and consequently on relatedness. LD is expected to break down over generations due to recombination events. However, although the estimated marker effects will need to be re-calibrated after a number of generations to avoid a decline in the prediction accuracy, regular collection of phenotypes and exposure to infection for all animals will not be required every generation, at least for several rounds of selection, and it will be possible to perform selection even in the absence of an epidemic. Further, high density genotyping becomes increasingly available at lower cost. The denser the markers are, the closer they will be to the QTLs and thus, the LD and the prediction accuracy will be more likely to be maintained across more generations.

Genomic selection in dairy cattle breeding presents further advantages over phenotypic selection. It improves the rate of genetic gain by obtaining sufficient

accuracy within a shorter generation interval, since the GEBVs can be calculated as soon as DNA samples are available. Hence, it allows differentiation between full-sibs, (i.e. prediction of the Mendelian segregation term), without the delay of phenotypic recording (Hayes et al. 2009; Daetwyler et al. 2008). Moreover, it reduces emphasis on any particular sire family and therefore, helps in controlling the rate of inbreeding (Daetwyler et al. 2008).

1.2.2 The Genomic selection opportunity

The hypothesis in the present study is that genetic selection for disease resistance in the light of genomic advances, may offer a complementary bTB control strategy, by reducing infection risks and hence contributing to a reduction in herd-level incidence as well as a reduction in the severity of the outbreak. Genomic selection is a new technology that allows to perform selection for disease resistance and overcome the limitations imposed by selection based on phenotypes or the knowledge of exact QTLs. Specifically, in the case of bTB, measuring resistance phenotypes under field conditions is challenging as data collection is opportunistic and relies on outbreaks. BTB is endemic but not present in all herds, thus only a small subset of herds each year can contribute data, while many herds have incomplete pedigree recording. Disease resistance traits often have low heritabilities which are further underestimated in field data due to the imperfect diagnostics (Bishop and Woolliams 2010). The phenotypes of interest are not directly observable and the indicator traits that are used i.e. the diagnostic tests, are imperfect in terms of their sensitivity and specificity. With genomic selection we can identify extreme (i.e. very resistant or very susceptible) animals and use GEBVs as predictors of disease

susceptibility without the requirement for exposure to infection and thus, genomic selection can be performed for animals without phenotypes, in the absence of an epidemic.

When genetic architecture allows, resistant animals can be selected through identifying the QTLs that control resistance to the disease, i.e. Marker Assisted Selection (MAS). One example of successful application of MAS for disease resistance is selection against Infections Pancreatic Necrosis (IPN) in the Atlantic salmon, where resistance to IPN was found to be controlled by a single QTL (http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=23913; Houston et al. 2010). However, the degree of success of MAS based on individual QTL associations depends on the proportion of the total genetic variation that they explain. Most disease resistance traits are expected to be complex traits, under polygenic control, i.e. affected by many genes with small individual effect (the theoretical model for this is known as the "infinitesimal model" where each variant explains only a very small proportion of the total variation) (Villanueva et al. 2005; Goddard 2009). In such cases, the ability to cope with large numbers of loci with small effects makes genomic selection more attractive than MAS. When using genomic selection and whole genome prediction for disease resistance traits, the QTLs do not need to be known, and SNP genotypes can be used to identify animals genetically predisposed to be more resistant e.g. to bTB infection. However, although widely implicated for production traits, the use of the genomic selection technology for disease resistance has been limited so far, although there are published examples of genomic selection for reduction of clinical mastitis in Norwegian Red cattle (Heringstad et al. 2003; Luan et al. 2009).

The aims of this PhD were to study the genetic architecture of resistance to bTB, i.e. to test whether single QTLs explain large amounts of variation following the IPN example, or if instead it is a complex trait following the mastitis example, and to explore the feasibility of genomic selection for livestock populations that will be more resistant to disease. Therefore, this Thesis presents seven chapters: Chapter 1 provides a general introduction. Chapter 2 demonstrates the feasibility of genomic selection for resistance to bTB. Chapter 3 examines a single-SNP approach capturing non-additive genetic variation. Chapter 4 progresses the idea of genomic selection by combining distantly related populations in a meta-analysis and exploring the genetic architecture of bTB resistance. To extend this analysis, Chapter 5 investigates the use of high density genotypes inferred by means of genotype imputation. Chapter 6 presents a quantitative genetic analysis of field SICCT data examining the potential impacts of selection for bTB resistance on the diagnostic test. Chapter 7 provides a general discussion of the overall Thesis.

Species		**h^2 ± SE**	**Scale**	**Ref.**
Red deer	*Cervus elaphus*	0.48 ± 0.10	lesion score	Mackintosh et al. 2000
Cattle	*Bos taurus*	0.18 ± 0.04	liability	Bermingham et al. 2009
Cattle	*Bos taurus*	0.18 ± 0.04	liability	Brotherstone et al. 2010
Cattle	*Bos taurus*	0.21 ± 0.06	observed	Bermingham et al. 2014
Cattle	*Bos taurus*	0.23 ± 0.06	observed	Tsairidou et al. 2014

Table 1. Heritability estimates reported in previous studies on bTB susceptibility.

Chapter 2

Genomic prediction for tuberculosis resistance in dairy cattle

2.1 Introduction

BTB eradication in the UK is impaired by limitations of the available diagnostic and control methods. Following exposure to *M. bovis* only a proportion of animals develop disease, implying variability among individuals in terms of their response to infection (Pollock et al. 2002). As described in Chapter 1 (section 1.2.2) selection of those animals that are genetically more resistant to bTB can be conducted using genetic markers and by applying the genomic selection technology. In contrast to traditional phenotypic selection, genomic selection allows us to obtain EBVs for animals without observed phenotypes and therefore, without exposure to infection to be required, at least for several rounds of selection.

Genomic selection utilises genomic EBVs estimated directly from SNP data, calculated as the sum of the effects of genetic markers (Single Nucleotide Polymorphisms, SNPs) across the genome. This is a conceptually different approach to single-SNP approaches. For example in GWA analysis (e.g. Bermingham et al. 2014), the objective is to identify causative variants, while in genomic selection entire breeding values are estimated using all markers in the genome. Several methods have been suggested for the calculation of EBVs either assuming that all markers explain the same amount of variance or incorporating prior knowledge on the distribution of the SNP effects (Meuwissen et al. 2001; Hayes et al. 2010; Moser et al. 2015). One method for estimating EBVs assuming the same amount of variance

is explained by all the markers, is Genomic BLUP (GBLUP) (Meuwissen et al. 2001; Daetwyler et al. 2010). Through the genomic prediction methodology, EBVs can be estimated by combining knowledge of marker effects estimated from a training population comprising both genotypes and phenotypes, with the genotypes of selection candidates. These can then be used as predictors of disease susceptibility for every animal.

Previous studies have confirmed the presence of potentially exploitable genetic variation in bTB susceptibility among dairy cattle (Brotherstone et al. 2010; Bermingham et al. 2012). The hypothesis in the present study is that genetic selection for disease resistance may offer a complementary bTB control strategy, by reducing infection risks and hence contributing to a reduction in herd-level incidence. The aim of this study was to estimate EBVs for bTB resistance using GBLUP, by utilising dense SNP chip data on UK dairy cattle and to test these genomic predictions in the absence of disease phenotype. This is the first step in the investigation of the feasibility of genomic selection for bTB resistance on the basis of predicted EBVs.

2.2 Materials and Methods

2.2.1 Animals

Phenotypic data for *1,151* cows from *165* dairy cattle herds in Northern Ireland were collected in a case-control study design, with a sample prevalence of *0.51* in the compiled dataset over two years of data collection (Bermingham et al. 2014). This prevalence is close to the cumulative incidence within Northern Ireland which was *0.66%* in 2012 and *0.51%* in 2013 (DARDNI 2013 Annual report).

Information available included bTB skin test data, as described below, the age of the cow on the day of the test, the year when the herd was tested, the season of the test, the reason for which the herd was tested and animal breed as nominated by the farmers. Animals were tested between August 2008 and September 2009, ranging in age from *1* to *11* years and with a mean of *4.8* years, either as part of the annual herd test, herd check tests or reactor herd tests (Abernethy et al. 2006). Most animals were assigned to be Holstein females, with a small number designated as Friesians *(n=164)*. A breakdown of data by these variables is given in Table 1.

The animal study was licensed by the Department of Health, Social Security and Public Safety for Northern Ireland (DHSSPSNI) under the UK Animals (Scientific Procedures) Act 1986 [ASPA], following a full Ethical Review Process by the Agri-Food & Biosciences Institute (AFBI) Veterinary Sciences Division (VSD) Ethical Review Committee. The study is covered by DHSSPSNI ASPA Project Licence (PPL-2638 'Host Genetic Factors in the Increasing Incidence of Bovine Tuberculosis'), and scientists and support staff working with live animals during the studies all hold DHSSPSNI ASPA Personal Licences.

2.2.2 Phenotype definitions

Cattle that showed a positive reaction to the Single Intradermal Comparative Cervical test (SICCT), that had bTB lesions confirmed by post-mortem examination of carcasses at slaughter and were confirmed as *M. bovis* positive by culture and molecular tests, were defined as cases (*592* animals). In this study a positive SICCT was defined as a skin test reaction to *M. bovis* antigens that after *72h* exceeds the reaction to *M. avium* antigens by *>4mm* according to the standard interpretation of

the test (Morrison et al. 2000; Bermingham et al. 2014). Controls were repeatedly SICCT negative, resident on the farm *>6* months (*559* animals), and were in herds where cases were observed (Bermingham et al. 2014). Controls were age-matched and preferentially selected from herds with higher disease prevalence in order to increase their probability of exposure to the pathogen (Bishop et al. 2012).

2.2.3 Genotyping

All individuals were genotyped for *727,252* SNPs using the BovineHD Illumina Bead Chip. Quality control parameters applied included a minimum Gentrain Call (GC) score of *0.60*, a minimum minor allele frequency of *0.05*, and a minimum call rate of *0.90* for all loci. Animals with a call rate *<90%* or a minimum GC score of *0.65* were excluded (Bermingham et al. 2014). The map of the SNP positions was based on the bovine genome assembly (*Bos taurus* UMD 3.0).

2.2.4 Structure exploration

Principal component analysis (PCA) was conducted in *R* (*R version 2.14*) to explore data structure by means of calculating the principal components. PCA allows identification of cryptic structure and unrecognised outlier groups representing subpopulations that are genetically distinct. PCA was conducted on the *1,151×1,151* identity-by-state (IBS) pairwise relatedness matrix using the BovineHD BeadChip genotypes (see section 2.2.6.2). Principal components were calculated as the eigenvectors of the IBS genomic matrix and the first principal component was plotted against the second principal component. PCA did not identify any sub-structure due to designated animal breed i.e. Holstein or Friesian, however, it

revealed the presence of two clusters, the main one, and a secondary smaller cluster comprising *40* individuals (Bermingham et al. 2014, Fig. S1). The hypothesis was that the observed structure might be due to the presence of Friesians in the dataset, however, none of the animals in the smaller cluster were described as Friesians. Thus, the structure observed in the data was not due to those breeds believed to have been sampled. Identification of the outliers showed that *39* of them originated from the same herd and further enquiries revealed that crossbreeding with beef cattle breeds may have taken place in this herd. Thus, to address the possibility of breed differences, these animals, along with one additional animal from a different herd that was also clustering with this group were deleted in some of the following analyses as described in the definition of datasets below.

2.2.5 Definition of datasets

Three slightly different datasets were used in this analysis. Firstly, the full dataset comprising all *1,151* individuals was used. Secondly, a reduced dataset was derived from the full dataset, removing the *40* individuals that were identified as outliers by the PCA and for which there was information that they could be crossbreds. This was done in order to address the hypothesis that the presence of beef cross-bred animals may introduce genetic structure to the population and hence alter prediction accuracy. Finally, the analysis was repeated using only animals designated as being Holsteins, after having removed the animals reported by the farmers as Friesians *(n=164)*. For each dataset the corresponding adjusted phenotypes were obtained, a new **G** matrix was calculated and the heritability was re-estimated (see section 2.2.6).

2.2.6 Calculating direct genomic estimated breeding values (EBV)

The aim of the analysis was to estimate the EBVs and then assess their predictive accuracy by cross validation. To conduct the cross validation, and to ensure that the sampling of phenotypes would not be biased by the fixed (non-genetic) effects, a two-step approach was followed to calculate the EBVs. Firstly, the data was pre-corrected for fixed effects, and then EBVs were obtained from the pre-adjusted data (de los Campos et al. 2012).

2.2.6.1 Pre-correction

An initial fixed effects model was used to obtain adjusted phenotypes, corrected for identifiable non-genetic factors. The fixed effects model included animal age, test year, season, test reason and breed as fixed effects, and was fitted using the ASReml package (Gilmour et al. 2009) as follows:

$$Y_{ijkmpq} = \mu + a_i + D_j + S_k + R_m + B_p + e_{ijkmpq} \quad (1)$$

where Y_{ijkmpq} represents the binary bTB status (*0*: control, *1*: case) of the q^{th} individual; μ is the overall mean; a_i is the age of the individual (*9* d.f.); D_j is the effect of the year of testing (*1* d.f.); S_k is the season of testing (*2* d.f.); R_m is the reason for which testing was initiated in the herd (*2* d.f.); B_p is the assigned breed of the individual (*1* d.f.) and e_{ijkmpq} is the residual error. Since all the animals were female, sex was not included in the fixed effects. The herd of origin was not included in the fixed effects as a consequence of choosing the controls to originate from herds of higher prevalence. The residual effects, which are independent of the fixed effects, were obtained and used as phenotypes for the subsequent analyses.

2.2.6.2 EBVs estimation

The genomic estimated breeding values were calculated for all individuals using the adjusted phenotypes from model (1). A random effects model was fitted in ASReml as follows:

$$y_i = m + u_i + e_i \quad (2)$$

where m is the overall mean, y_i is the residual effect for the i^{th} individual as calculated from model (1), u_i is the genomic estimated breeding value with $\mathbf{u} \sim$ MVN $(0, \mathbf{G}\sigma_a^2)$ and e_i is its residual value with $\mathbf{e} \sim$ MVN $(0, \mathbf{I}\sigma_e^2)$. As pedigree relationships were unknown in this population, genetic similarities between animals were described using the marker-based IBS genomic relationship (**G**) matrix which has the following elements:

$$f_{ij} = \frac{2}{n}\sum_{k=1}^{n}\frac{(x_{ik}-p_k)(x_{jk}-p_k)}{p_k(1-p_k)}, (i \neq j)$$

$$f_{ii} = 1+\frac{1}{n}\sum_{k=1}^{n}\frac{Obs(\hom)_{ik}-E(\hom)_k}{1-E(\hom)_k}, (i = j)$$

where x_{ik} (x_{jk}) is the genotype of the i^{th} (j^{th}) animal defined as *0*, *½*, or *1* from an arbitrary reference allele at the k^{th} SNP, n the total genomic SNPs, and p_k is the frequency of the reference allele at the k^{th} SNP. $Obs(hom)_{ik}$ is the observed homozygosity (1 or 0) for the i^{th} animal at the k^{th} SNP and $E(hom)_k$ is the expected homozygosity for the k^{th} SNP calculated as $E(hom)_k = (1 - 2p_k(1-p_k))$ (Uemoto et al. 2013). To construct **G**, SNPs found only in the homozygote state in the sample and those found on the X chromosome were removed (*601,280* SNPs were finally

retained in the analysis). In ASReml the provision of the inverse **G** matrix is required which was calculated in *R*.

2.2.7 Heritability estimation

For the purpose of estimating the heritability of bTB resistance from the full dataset, the fixed and random effects were fitted simultaneously in a mixed model in ASReml, where all the fixed effects from model (1) were fitted as before and the relationship information from the **G** matrix was incorporated as a random effect with distributional assumptions as in model (1).

2.2.8 Cross validation

Genomic prediction accuracy can be assessed through cross validation, a non-parametric method that allows assessment of the predictive ability of the EBVs. By partitioning the data into a training set and a validation set, EBVs can be predicted for the validation set without reference to their phenotypic information. Prediction accuracy can then be calculated by correlating the predicted breeding values and the observed phenotypes, corrected for trait heritability (Legarra et al. 2008). A five-fold cross validation was conducted as follows.

Firstly, to create the training set in each of the three datasets the individuals were partitioned into five random groups of near-equal size, with the randomization performed separately within the case and control sub-populations. Phenotypes were then masked for each subset in turn, creating five datasets (or folds) in which four-fifths of the animals had a phenotype (training-set, y_1), and one-fifth had no phenotype (validation-set, y_2).

Secondly, using the GBLUP model (2) predicted EBVs were calculated for each validation-set in turn based on the **G** matrix alone and conditional on the phenotypic information recorded on the training-set animals, ($\hat{y}_2|y_1$) (Legarra et al. 2008; Luan et al. 2009; Daetwyler et al. 2013).

For each of the five test-sets the correlation between the cross-validated predicted EBVs *(ŷ)* and the adjusted phenotypes *(y)*, i.e. *r(y,ŷ),* was calculated. The expected accuracy *(r(g,ĝ))* between the breeding value of an individual *(g)* and its estimate *(ĝ)*, was derived from the correlation as $E[r(g,\hat{g})] \approx r(y, \hat{y})/h$, where h is the square root of the heritability (Luan et al. 2009). The accuracy for each test set was calculated using the heritability obtained for each corresponding cross validation fold and then the average accuracy across all the individuals was obtained.

In order to reduce random sampling effects and assess the sampling properties of the accuracy, the cross validation analysis as described above was replicated six times, where for each replicate a new randomisation was performed so that the individuals comprising each of the groups were different. Finally the average accuracy across all six replications with its empirical *95%* confidence interval was obtained, where the confidence interval was calculated from a one sample t-test (*5* d.f.) for the six accuracy values obtained from the six replications.

2.2.9 Assessing predictive ability using ROC curves

Genomic predictions can be further assessed through the properties of the Receiver Operator Characteristic (ROC) curves and the corresponding area-under-the-ROC-curve (AUC). A ROC curve is the plot of the probability of a positive test result given that the individual is diseased (sensitivity) versus the probability of a

positive test result given that the individual is healthy (1-specificity), for all successive thresholds (Metz 1978). The AUC is a measure of the performance of the predictor, i.e. the AUC is the probability of correctly identifying the case in a pair of infected and healthy individuals (Janssens et al. 2006). Using the *R* package, the predicted EBVs for each of the omitted (validation) groups from the cross validation procedure and the binary phenotype for all the *1,151* individuals were used to calculate the ROC curves, along with their corresponding AUC values, for each of the six randomisations for the full dataset.

2.2.10 Theoretical expectations

2.2.10.1 The maximum AUC value (AUC_{max})

Insight into the information obtained by calculating the ROC curves and their corresponding AUC can be gained by considering these values relative to the theoretical maximum AUC value that could be obtained given the characteristics of the trait and the population under study. There is a maximum AUC value (AUC_{max}) that would be achieved if the test classifier was a perfect predictor of genetic risk (Wray et al. 2010). This maximum varies for each disease, since it depends on the disease prevalence *(q)* and the heritability of the trait on the underlying liability scale *(h_L^2)*. h_L^2 can be estimated from the approximation $h_o^2 \sim h_L^2 q^2 i_q^2 [q(1-q)]^{-1}$ as introduced by Robertson and Lerner (1949), where h_o^2 is the heritability on the observed scale, q is the disease prevalence in the sample and i_q is the mean in standard deviation units of the upper proportion q of the population, assuming a normal distribution. The online calculator provided by Wray et al. (2010) was used to obtain expected values for AUC_{max} and AUC_{half}, which is defined as the AUC

expected from a genomic profile that accounts for only a half of the known genetic variance (i.e. a reliability of *0.5*). These values can be used as a basis of comparison for the actual AUC values obtained in the present study.

2.2.10.2 Prediction accuracy

Daetwyler at al. (2010) presented a formula for estimating the expected GBLUP accuracy:

$$r(\mathrm{g}, \hat{g}) = \sqrt{\left[N_P h^2 / \left(N_P h^2 + M_e \right) \right]} \qquad (3)$$

where N_P is the number of individuals in the training population, h^2 is the heritability on the observed scale, and M_e is defined as the number of independent genome segments. The formula for M_e is given by Meuwissen et al. (2009) where Σ represents the sum across chromosomes:

$$M_e = \sum_i 2N_e L_i / \ln\left(4N_e L_i\right) \qquad (4)$$

M_e depends on the chromosome length in Morgans L_i (Lee et al. 2011) and on the effective population size N_e. Formulae (4) and (5) were applied to different putative effective population sizes for this sample of animals in order to obtain estimates for the number of independent chromosome segments and the expected corresponding prediction accuracy, for the full dataset and the dataset without the Friesians.

2.3 Results

2.3.1 Calculation of EBVs and genomic prediction accuracy

The GBLUP analysis gave an estimate for the heritability of bTB susceptibility of $h^2 = 0.23 \pm 0.06$ on the observed scale and $h_L^2 = 0.34$ on the liability scale for the full data set, $h^2 = 0.23 \pm 0.07$ and $h_L^2 = 0.34$ for the dataset after removing the *40* individuals identified as a distinct sub-population from the PCA, and $h^2 = 0.21 \pm 0.07$ with $h_L^2 = 0.34$ for the reduced data set with the Friesian individuals excluded.

Table 2 shows the correlations between the adjusted phenotypes and the predicted EBVs, the corresponding heritability estimates and accuracy values with their standard deviations obtained as averages across the five cross validation groups for each of the six replications. More detailed information is presented in Tables 3, 4, and 5. Accuracies of *0.33 (95% C.I.: 0.26, 0.40), 0.33 (95% C.I.: 0.28, 0.37)*, and *0.36 (95% C.I.: 0.33, 0.38)* were obtained for the three datasets, respectively. Analysis after removing the *164* animals designated as Friesians provided more homogenous results across the cross validation repeats (Fig. 1). As it will be discussed below, the values obtained are in line with theoretical expectations given the size of the dataset.

Further, for the full data and the dataset after removing the Friesians, for each of the cross validation folds and across the six replications, the observed phenotypes were regressed on the predicted EBVs (Tables 6, 7, and 8). For the full dataset, these values were close to the theoretical value of *1.0* indicating that the predicted EBVs

were unbiased. After removing the Friesians the *SD* reduced, however, although the regression coefficients might indicate some bias, the number of records in this analysis was smaller.

2.3.2 ROC curves and AUC values

ROC curves, showing the utility of EBVs as predictors of the binary phenotype, are shown in Figure 2. The ROC curves result from plotting for all successive decision thresholds, the true positive fraction versus the false positive fraction which can be defined as the conditional probability of a positive test given the presence of disease, and the conditional probability of a positive test given the absence of disease respectively, and thus, are independent from the decision threshold and the prevalence of the disease (Metz 1978). In the context of estimated EBVs, the ROC curves are the plots of the conditional probability of EBVs predicting a case given a diseased true phenotype versus the probability of EBVs predicting a case given a healthy true phenotype. Examples of individual ROC curves for each of the five cross validation test sets within one cross validation run are shown in Figure 3. In these ROC curve plots, the comparison of interest is with the outcome that would be expected by chance (diagonal line of no discrimination). The curves for all randomisations lie above this diagonal line. Therefore, for the population under study the use of genotypes provides information in the prediction of disease state, i.e. the markers help to predict resistance. The AUC values were *0.56, 0.59, 0.58, 0.57, 0.57* and *0.59* for the six different randomizations applied in dataset 1 (Fig. 2). Hence, there was a probability close to *0.58* of correctly classifying a case cow and a control cow based on SNP genotype alone using these data.

2.3.3 Theoretical expectations

2.3.3.1 AUC values

For the data-set in the present study the disease prevalence was *0.51* (*592* cases out of *1,151* animals in total) and h_L^2 was estimated to be *0.34* for a heritability on the observed scale of *0.23*. For a prevalence $p = 0.5$, the selection intensity (i_q) would be *0.798* (Falconer et al. 1997). An $AUC_{max} = 0.77$ and $AUC_{half} = 0.69$, can then be obtained using the online calculator provided by Wray et al. (2010). Therefore, the maximum achievable accuracy in this dataset would be *0.77*. Our AUC value of *0.58* is somewhat less than AUC_{half}, i.e. this is consistent with the accuracy value which also was notably less than *0.7* which is the accuracy corresponding to AUC_{half}.

2.3.3.2 Prediction accuracy

Expected accuracies of the genomic predictions are shown in Table 9. This approach combines the effective population size, with the accuracy and the heritability. With N_P being the average number of individuals in the training population *(920.8* i.e. *4/5* of full dataset*)*, and h^2 the heritability on the observed scale *(0.23)*, the number of independent chromosome segments M_e was calculated for different values of effective population size (Table 9). If ΔF_g is the rate of inbreeding per generation, then for a rate of inbreeding per year $\Delta F_y = 0.0017$ (Kearney et al. 2004) and a five years generation interval for dairy cattle, $\Delta F_g \approx 0.01$ and thus, a suggestive value for the effective population size would be $N_e \approx 50$. Using formulae 4 and 5, with $N_e \approx 50$ and $M_e = 639.79$, the expected accuracy would be $r(g,\hat{g}) = 0.50$. Reversing the calculations, an expected accuracy of $r(g,\hat{g}) = 0.34$, gives an

effective population size of $Ne \approx 150$. This value may not be an unreasonable value for the Holstein-Friesian cows in this sample, given that the population under study is a sample derived as a random selection of non-pedigree dairy cattle and hence possibly not as highly selected as cattle recorded in pedigree databases. Thus, there are likely to be Friesian cows in the dataset along with the possibility of a small number of crossbred animals.

For the dataset with the animals designated as Friesians excluded, the expected accuracies were slightly lower, and the observed accuracy was consistent with an effective population size of ca. *100* individuals, however, the heritability was also lower compared to the full dataset (Table 9).

2.4 Discussion

This study provides evidence that genomic selection for bTB resistance is potentially feasible in populations where phenotypic information in unavailable for selection candidates, and even when no pedigree is available. Genomic selection can be considered as a two-step procedure. Initially, on a reference population with both phenotypic and genotypic information, EBVs can be calculated as the genome-wide sum of marker effects (Luan et al. 2009). Then, for selection candidates the EBVs can be predicted without the need for phenotypes, since the marker effects have already been calculated in a relevant reference population (Habier et al. 2007). With this design, the results of the present study are important in the context of bTB control. Predicting EBVs in the absence of phenotypes is highly beneficial in the case of bTB, as collection of appropriate phenotypic information requires that a population undergoes an epidemic and that all animals (including controls) are

exposed to the pathogen (Bishop et al. 2010). These conditions can only be met for a subset of animals in the national population and will become increasingly difficult to satisfy as disease prevalence decreases in later stages of eradication programmes.

The predictive accuracy of the EBVs is at levels that justify further studies on larger populations in order to obtain predictions that could be used in evaluation of selection candidates for their bTB resistance. In order to obtain an accuracy of *0.7*, the theoretically required number of animals needed in the training population can be calculated by rearranging formula (3). Given a heritability of *0.23* and with N_e *= 50, ~2,670* individuals would be needed in the training population comprising both cases and controls with both phenotypes and genotypes. But if the N_e were to increase to *100*, the size of the required training population would increase to *~4,747* individuals, as might be expected, and if N_e was *150* then N_P would be *~6,685*. However, if for example the N_e was *100* but we targeted a prediction accuracy of *0.5*, then the size of the training population needed would reduce to *~1,647*. Although in our study, the size of the training population *(920.8)* was somewhat smaller, the outcomes of the analyses suggest that genomic selection is potentially feasible. However, implementation of genomic selection should wait until we have a greater number of individuals in the training population, to enable us to achieve higher accuracy.

2.4.1 Estimated heritability

The data set of UK dairy cattle analysed in this study through the GBLUP approach, provided a heritability estimate of *0.23* (*0.34* on the liability scale) for the trait of tuberculosis resistance. This value indicates stronger evidence for genetic

variation than previous estimates (Bermingham et al. 2009; Brotherstone et al. 2010), and our estimate is lower than the value reported for deliberately challenged red deer (Mackintosh et al. 2000). However, direct comparison between these studies, some with pedigree information and some without, should be undertaken with caution. Health traits often have low reported heritabilities and the estimates obtained are influenced by the experimental design and data recording strategy, the imperfect diagnostic tests, and the incomplete exposure to the pathogens (Bishop and Woolliams 2010; Bishop et al. 2012). However, the common conclusion of intermediate heritability of tuberculosis resistance makes genomic selection for bTB resistance an appealing approach to assist in bTB control.

2.4.2 ROC curves properties

A ROC curve is a representation of the different combinations of sensitivity and specificity for successive thresholds between a positive and a negative test result. Although the ROC curves and their AUCs based on genotypic information in this study show only a modest increase in the probability of correctly classifying cases or controls compared to random expectations, these values should also be considered relative to the AUC_{max} (Wray et al. 2010). This represents an upper limit of predictive ability given the properties of the dataset and the trait under study, assuming that the classifier (i.e. the EBVs) were a perfect predictor of genetic risk. Since AUC_{max} depends on disease prevalence and trait heritability, the authors argue that prediction accuracy should be preferred as a measure for evaluating genomic predictions (Wray et al. 2007).

2.4.3 Cross validation prediction accuracy

Random error due to sampling effects was minimized by averaging the accuracies across several replications with different randomizations so that the individuals comprising each of the five groups were different each time. The differences observed between the randomizations indicate that even with ca. *1000* individuals, random sampling effects still contribute significantly to the cross validation outcomes (Fig. 1). Conducting more randomisations was preferred to increasing the number of groups i.e. cross validation folds, because the test set would be reduced, thereby increasing variability across the cross validation folds through samples.

When the full dataset was used, the accuracy obtained was consistent with the theoretical accuracy obtained using the formula by Daetwyler et al. (2008) for an effective population size of $N_e = 150$, given the properties of the dataset (i.e. sample size and trait heritability). This N_e value is somewhat higher than that often suggested for the Holstein cattle population (c.f. N_e ca. *50*; McParland et al. 2007), but it is possible that this is representative of the sample population as it may have been inflated due to the structure present in the dataset revealed by PCA, and also from the designation of several individuals as Friesian. Both factors would increase LD and consequently the apparent N_e. Further, as indicated in 2.3.3.2, the population under study is not a pedigree or a highly selected population, with the animals included in the study sampled from random commercial farms.

It should be noted that results from the different variations of the datasets used were coherent across the analyses. When the cows designated as Friesians were

removed, in addition to giving slightly increased accuracy, the dataset behaved more consistently across replicates (Fig. 1), and the corresponding implied N_e was reduced (N_e ca. *100*) but with a lower heritability. This small increase in the accuracy was despite the fact that the dataset was smaller; presumably reflecting a more uniform population with linkage disequilibrium extending across longer chromosomal regions. Removing the PCA outliers had little impact on the prediction accuracy, however the animals removed (*14* cases and *26* controls) may have been too few to greatly influence the results.

2.4.4 Prediction accuracy and epidemic properties

2.4.4.1 Imperfect diagnostics

Some insight into how the epidemic properties might have an impact on the prediction accuracy can be gained if we consider the diagnostic test and that it might be imperfect in terms of its sensitivity and specificity. Although the qualities of the diagnostic test might not be strictly considered as a property of the epidemic itself, analyses gain information on the epidemic in part through the diagnostic test and the control measures that are adopted incorporating their use. Therefore, the diagnostics and the epidemic properties become interdependent. With an imperfect diagnostic test the estimated prediction accuracy in the sample under study will be different from the true accuracy that would be achieved in the population under a genomic selection scenario.

The heritability on the observed scale is calculated from the sample and thus, might be different from the true heritability of the trait in the population. The observed heritability depends on the disease prevalence in the sample which depends

on the sensitivity and specificity of the diagnostic test and can be calculated from the true prevalence as: $p'=(1-Sp)+(Sp+Se-1)p$ (Bishop and Woolliams 2010; Bishop et al. 2012). The heritability on the liability scale is calculated from the heritability on the observed scale following $h_L^2 p^2 i_p^2 [p(1-p)]^{-1}$ (Robertson and Lerner 1949), where p is the sample prevalence (p'). With imperfect diagnostics the heritability on the liability scale is underestimated and specifically when the true prevalence is less than *0.5*, it has been shown that imperfect specificity has a greater impact on underestimating the heritability on the liability scale (Bishop and Woolliams 2010). For bTB however, the diagnostic comparative tuberculin skin test has very good specificity, thus the impact of the imperfect sensitivity on the heritability estimation would be expected to be less detrimental. Moreover, in the present study, the cases were confirmed through post mortem examination and laboratory confirmation and thus the achieved specificity was very good. Therefore, and although still sensitive to changes of the sample size, the estimated accuracy in the present study is not influenced by the diagnostic test sensitivity and specificity.

2.4.4.2 The case-control design

The study design can have an impact on the prediction accuracy through the estimated marker effects. When using the case-control design, selection intensity is applied on the groups of cases and controls, increasing the between-group genetic variance in the sample, and this can influence the SNP effects as follows: (a) The overall precision of estimated SNP effects is improved due to increasing the between-group variance in the case-control sample compared to a random sample from the general population, and will offer an improvement in accuracy. (b) The power of estimating the effects of rare variants that might be linked to susceptibility

is improved, through increasing their frequency in the sample by increasing the fraction of cases. (c) Some bias is introduced to the SNP effects that would be observed in the general population due to reducing the overall frequency of heterozygotes and shifting the allele frequencies towards intermediate values. (d) When using a predictor of susceptibility on the observed scale, the accuracy of predicting phenotypes estimated in the case-control sample will be greater than the accuracy that would be achieved in the population. However, the heritability on the observed scale also decreases as prevalence decreases when going from the case/control sample with sample prevalence of *~0.5,* to the real population with lower true prevalence. When in the present study, the prediction accuracy for the EBVs was scaled by the observed heritability ($E[r(g,\hat{g})] \approx r(y, \hat{y})/h$), the change of the prevalence when moving from the case-control sample to the population, was not expected to have a detrimental impact on the accuracy of the EBVs due to these compensating changes (Woolliams J., personal communication, September 7, 2015).

2.4.5 Conclusion

Our results demonstrate that genomic selection for bTB resistance is feasible in principle even in populations with no pedigree recording, and it can be applied to animals lacking bTB phenotypes. Genomic prediction accuracies in the present study reflected the expected values given the size of the dataset and the LD structure of the genotypes which is a function of the effective population size for the Holstein breed. Access to a greater number of animal phenotypes, thereby creating larger training sets, is expected to improve prediction accuracies and open up opportunities for implementation.

	Year		Season			Test reason		
	2008	**2009**	**Winter**	**Spring**	**Autumn**	**Annual**	**Herd check**	**Reactor herd**
Cases	359	233	309	115	168	155	231	206
Controls	384	175	253	96	210	124	251	184
Totals	743	408	562	211	378	279	482	390

Table 1. The number of animals in the dataset, classified by year of test, season of test and reason for test.

	Full Dataset				Excluding minor cluster				Excluding Friesians			
	r(y,ŷ)	h^2	r(g,ĝ)	SD	r(y,ŷ)	h^2	r(g,ĝ)	SD	r(y,ŷ)	h^2	r(g,ĝ)	SD
Run 1	0.10	0.21	0.22	0.12	0.13	0.21	0.29	0.05	0.13	0.18	0.34	0.22
Run 2	0.15	0.19	0.36	0.08	0.15	0.20	0.35	0.10	0.15	0.17	0.38	0.10
Run 3	0.15	0.20	0.34	0.14	0.12	0.21	0.29	0.17	0.14	0.18	0.35	0.18
Run 4	0.14	0.20	0.33	0.17	0.14	0.20	0.34	0.25	0.15	0.17	0.37	0.16
Run 5	0.13	0.20	0.31	0.11	0.16	0.19	0.40	0.21	0.15	0.17	0.37	0.18
Run 6	0.17	0.19	0.42	0.18	0.12	0.21	0.28	0.19	0.13	0.18	0.32	0.07
Average	0.14	0.20	**0.33**	**0.07**	0.14	0.21	**0.33**	**0.05**	0.14	0.18	**0.36**	**0.02**

Table 2. Correlations between adjusted phenotypes and predicted EBVs (*r(y,ŷ)*), heritabilities (h^2) and prediction accuracies (*r(g,ĝ)*), for each of the six replicates of full cross-validation. SD represents the sampling standard deviation among folds within a run, and for the average represents the SD over the mean accuracy values for each replicate. In this table are shown the parameter values for each of the cross validation runs and the averages across all replications for the full data set, the reduced dataset after having removed the animals clustering separately in the PCA, and for the dataset without the animals designated as Friesians.

	Run 1				Run 2				Run 3				Run 4				Run 5				Run 6			
	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$
Group 1	0.08	0.23	0.08	0.17	0.13	0.20	0.08	0.30	0.20	0.16	0.08	0.50	0.11	0.22	0.08	0.24	0.07	0.22	0.08	0.15	0.19	0.18	0.08	0.45
Group 2	0.08	0.22	0.08	0.16	0.17	0.19	0.08	0.40	0.08	0.20	0.07	0.17	0.12	0.25	0.08	0.24	0.20	0.25	0.08	0.40	0.11	0.23	0.08	0.23
Group 3	0.15	0.17	0.07	0.33	0.18	0.17	0.08	0.44	0.12	0.25	0.08	0.25	0.22	0.13	0.07	0.62	0.14	0.13	0.07	0.39	0.24	0.12	0.07	0.70
Group 4	0.16	0.18	0.07	0.36	0.12	0.22	0.08	0.24	0.18	0.17	0.08	0.44	0.10	0.19	0.07	0.23	0.10	0.19	0.07	0.23	0.13	0.21	0.08	0.29
Group 5	0.05	0.25	0.08	0.09	0.17	0.17	0.07	0.41	0.16	0.20	0.08	0.36	0.15	0.19	0.07	0.33	0.16	0.19	0.07	0.36	0.18	0.20	0.08	0.41
Average	0.10	0.21		**0.22**	0.15	0.19		**0.36**	0.15	0.20		**0.34**	0.14	0.20		**0.33**	0.13	0.20		**0.31**	0.17	0.19		**0.42**

Table 3. For the data set including all the individuals, the correlation, heritability with its standard error and corresponding prediction accuracy for each of the five test-groups from the Cross Validation procedure are presented for the six different randomization replications.

	Run 1				Run 2				Run 3				Run 4				Run 5				Run 6			
	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g, \hat{g})$	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g, \hat{g})$	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g, \hat{g})$	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g, \hat{g})$	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g, \hat{g})$	$r(\hat{y}_2, y_2)$	h^2 and SE		$r(g, \hat{g})$
Group 1	0.15	0.18	0.08	0.36	0.13	0.26	0.09	0.25	0.06	0.27	0.08	0.12	0.05	0.25	0.08	0.10	0.17	0.22	0.08	0.36	0.08	0.25	0.08	0.15
Group 2	0.13	0.20	0.08	0.29	0.18	0.15	0.08	0.45	0.13	0.17	0.08	0.31	0.20	0.15	0.08	0.50	0.12	0.22	0.08	0.26	0.15	0.23	0.08	0.32
Group 3	0.14	0.20	0.08	0.31	0.20	0.18	0.08	0.46	0.17	0.18	0.08	0.40	0.25	0.13	0.08	0.69	0.23	0.14	0.08	0.62	0.03	0.26	0.08	0.06
Group 4	0.12	0.22	0.08	0.26	0.15	0.21	0.08	0.32	0.19	0.15	0.08	0.50	0.14	0.20	0.08	0.31	0.08	0.28	0.09	0.15	0.14	0.19	0.08	0.32
Group 5	0.11	0.23	0.08	0.24	0.13	0.21	0.08	0.28	0.06	0.30	0.09	0.10	0.06	0.29	0.09	0.11	0.21	0.11	0.07	0.62	0.21	0.14	0.08	0.56
Average	0.13	0.21		**0.29**	0.15	0.20		**0.35**	0.12	0.21		**0.29**	0.14	0.20		**0.34**	0.16	0.19		**0.40**	0.12	0.21		**0.28**

Table 4. For the data set in which animals clustering separately in the PCA were removed, the correlation, heritability with its standard error and corresponding prediction accuracy for each of the five test-groups from the Cross Validation procedure are presented for the six different randomisation replications.

	Run 1				Run 2				Run 3				Run 4				Run 5				Run 6			
	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2\ y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2\ y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$	$r(\hat{y}_2,y_2)$	h^2 and SE		$r(g,\hat{g})$
Group 1	0.17	0.14	0.08	0.44	0.18	0.16	0.08	0.45	0.17	0.13	0.07	0.47	0.15	0.19	0.08	0.33	0.04	0.26	0.09	0.08	0.15	0.16	0.08	0.38
Group 2	0.23	0.13	0.08	0.63	0.18	0.13	0.07	0.51	0.15	0.19	0.08	0.34	0.07	0.20	0.08	0.17	0.13	0.17	0.08	0.31	0.14	0.16	0.08	0.35
Group 3	0.09	0.18	0.08	0.21	0.15	0.18	0.08	0.36	0.13	0.22	0.09	0.28	0.17	0.14	0.08	0.47	0.18	0.14	0.08	0.49	0.12	0.17	0.08	0.30
Group 4	0.15	0.18	0.08	0.36	0.14	0.16	0.08	0.33	0.05	0.26	0.09	0.09	0.14	0.19	0.09	0.32	0.20	0.15	0.08	0.52	0.16	0.19	0.09	0.38
Group 5	0.02	0.25	0.09	0.04	0.12	0.23	0.09	0.25	0.19	0.12	0.08	0.57	0.21	0.13	0.08	0.58	0.17	0.16	0.08	0.44	0.10	0.22	0.09	0.21
Average	0.13	0.18		**0.34**	0.15	0.17		**0.38**	0.14	0.18		**0.35**	0.15	0.17		**0.37**	0.15	0.17		**0.37**	0.13	0.18		**0.32**

Table 5. For the dataset when the *164* animals designated as Friesians were removed, the correlation, heritability with standard errors and corresponding prediction accuracy for each of the five test groups in the Cross Validation procedure resulting from the six randomisation replications. The data for the remaining *987* animals were re-randomised to training and test sets, which were *~790* and *~198* respectively. In the initial fixed effects model breed was removed from the fixed effects and a new **G** matrix calculated only for the Holsteins was used.

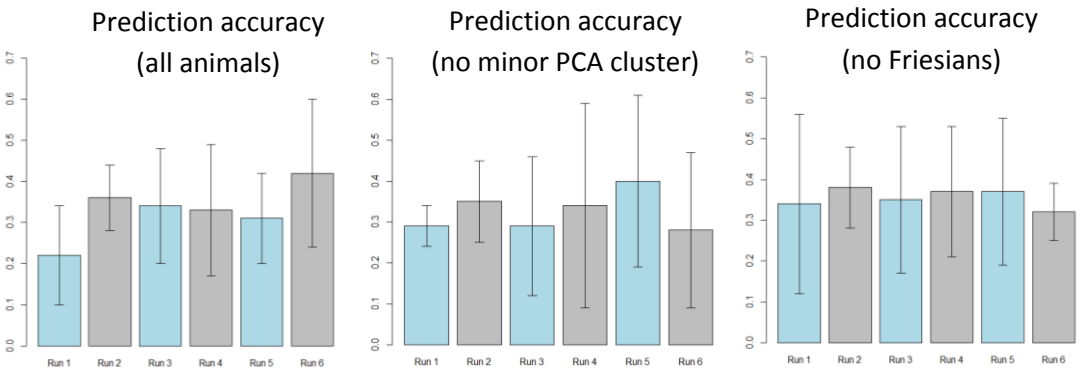


Figure 1. Mean prediction accuracy values with corresponding standard deviations for each of the six randomisation replications in the cross validation procedure for the data set including all the individuals, the data set in which animals clustering separately in the PCA were removed, and for the dataset when the *164* animals designated as Friesians were removed.

	Regression coefficient	SD	Regression coefficient	SD
Run 1	0.74	0.41	1.17	0.87
Run 2	1.14	0.27	1.31	0.45
Run 3	1.08	0.43	1.22	0.75
Run 4	1.16	0.78	1.26	0.55
Run 5	1.16	0.78	1.31	0.71
Run 6	1.42	0.75	1.06	0.24
Average	1.11	0.22	1.22	0.10

Table 6. For the regression of phenotypes on predicted EBVs, average regression coefficients among test sets for each of the cross validation runs and the average across all replications, with the corresponding standard deviations where the SD for the average is the SD of the means presented for the six runs. Left part of the table: full data set, right part: dataset from which the Friesians were excluded.

	Group 1		Group 2		Group 3		Group 4		Group 5			
	a	**b**	**a**	**b**	**a**	**b**	**a**	**b**	**a**	**b**	**Intercept SD**	**Regr coef SD**
Run 1	-0.002	0.580	-0.014	0.502	0.002	1.099	0.017	1.237	0.006	0.279	0.011	0.410
Run 2	-0.014	0.937	0.004	1.345	0.009	1.302	-0.004	0.766	0.004	1.333	0.009	0.268
Run 3	0.023	1.474	-0.006	0.511	-0.007	0.783	-0.001	1.506	-0.002	1.108	0.012	0.433
Run 4	0.012	0.834	-0.011	0.656	-0.003	2.530	-0.007	0.701	0.011	1.063	0.010	0.784
Run 5	0.012	0.834	-0.011	0.656	-0.003	2.530	-0.007	0.701	0.011	1.064	0.010	0.784
Run 6	0.002	1.491	0.006	0.636	0.004	2.598	-0.006	0.950	-0.005	1.400	0.005	0.746

Table 7. For the regression analysis on the full dataset, intercept (a) and regression coefficients (b) for the regression of adjusted phenotypes (observed) on cross-validated EBVs (predicted), for each cross validation fold across the six replication runs, with corresponding standard deviations.

	Group 1		Group 2		Group 3		Group 4		Group 5			
	a	b	a	b	a	b	a	b	a	b	Intercept SD	Regr coef SD
Run 1	0.001	1.371	0.005	2.429	0.005	0.635	-0.006	1.325	0.002	0.114	0.004	0.873
Run 2	0.000	1.629	-0.008	1.918	0.007	1.185	-0.001	1.001	-0.002	0.834	0.005	0.450
Run 3	-0.001	1.747	0.010	1.093	0.007	0.860	-0.001	0.261	-0.017	2.157	0.011	0.745
Run 4	0.002	1.077	0.004	0.542	0.000	1.754	-0.009	1.055	0.010	1.870	0.007	0.549
Run 5	-0.002	0.228	0.002	1.095	-0.008	1.769	0.005	2.071	0.007	1.385	0.006	0.709
Run 6	-0.006	1.229	0.005	1.200	-0.006	1.043	0.000	1.176	0.010	0.651	0.007	0.240

Table 8. For the regression analysis on the datset without the Friesians, intercept (a) and regression coefficients (b) for the regression of adjusted phenotypes (observed) on cross validated EBVs (predicted), for each cross validation fold across the six replication runs, with corresponding standard deviations.

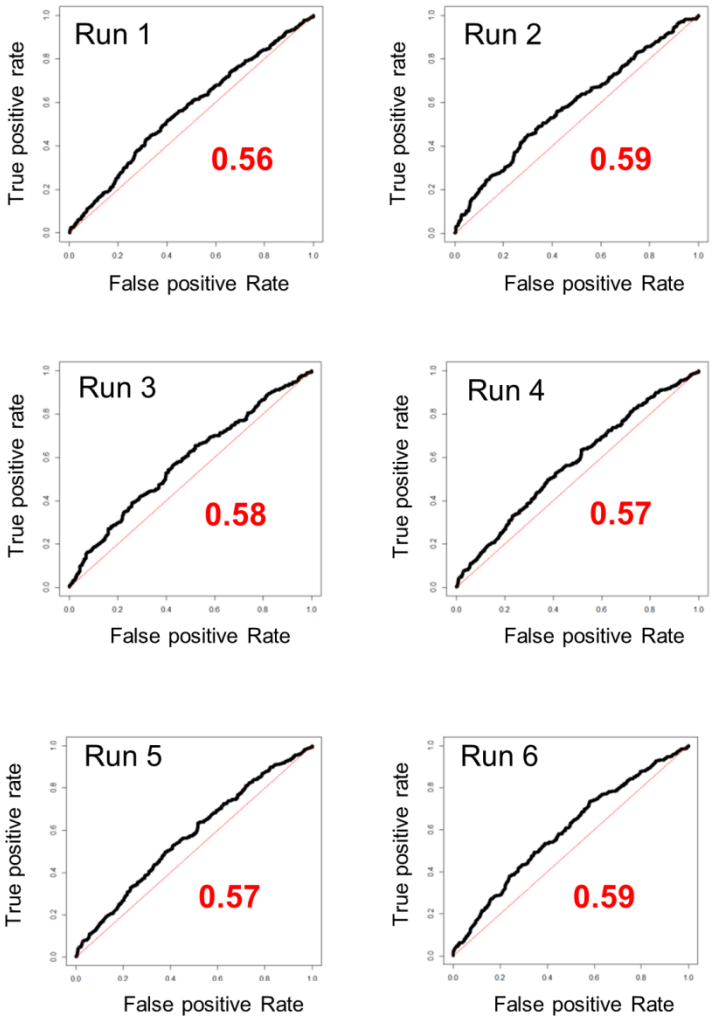


Figure 2. ROC curves (a plot of the true positive rate, i.e. the sensitivity, against the false positive rate, i.e. *1*-specificity), and the corresponding AUC (the probability of correctly assigning an individual as diseased or as healthy on the basis of its genotype alone) for the six randomisation runs for the full dataset.

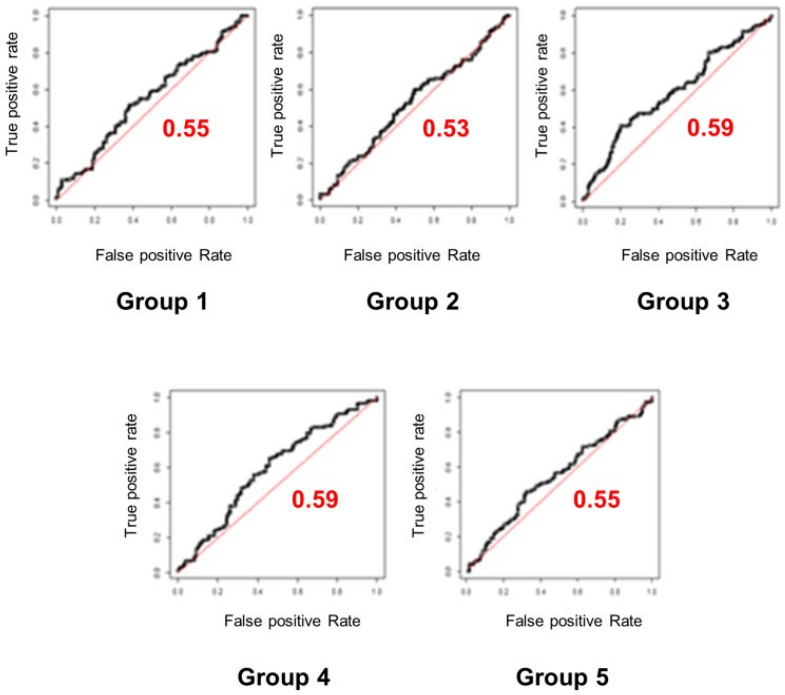


Figure 3. For the full data set, the ROC curves for each of the five cross validation test groups are presented for the first randomisation.

	Full dataset (N_P = 920.8 and h^2 = 0.23)		Excluding Friesians (N_P = 789.6 and h^2 = 0.21)	
Assumed N_e	**$\sum M_e$**	**r(g,ĝ)**	**$\sum M_e$**	**r(g,ĝ)**
50	639.79	0.50	639.79	0.45
100	1136.53	0.40	1136.53	0.36
150	1600.18	0.34	1600.18	0.31

Table 9. Training population size *(N_p)*, heritability, number of independent genome segments *($\sum M_e$)* and corresponding expected accuracy *(r(g,ĝ))* for different assumed effective population sizes. Left part of the table: full data set, right part: dataset from which the Friesians were excluded.

Chapter 3

An analysis of heterozygote advantage for tuberculosis resistance in dairy cattle

3.1 Introduction

Animals that are more resistant to bTB can be selected using genetic markers likely to be in linkage disequilibrium with a QTL. In the previous chapter it was demonstrated the feasibility of genomic prediction for bTB resistance exploiting information across the entire genome. An alternative approach for selecting more resistant individuals is through identifying the individual QTL(s) that are associated with the trait under study. This knowledge can then be employed in Marker Assisted Selection (MAS).

There are previous examples of successful implementation of selection for disease resistance based on specific markers. Studies on Infectious Pancreatic Necrosis (IPN) in the Atlantic salmon revealed the presence of a single QTL, explaining almost all the genetic variation (Houston et al. 2010), and the results of those studies have been successfully used in MAS (http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=23913). Plans for selection for resistance to Scrapie in sheep were based on the PrP locus (http://adlib.everysite.co.uk/resources/000/054/063/NSP_english.pdf). Selection against malignant hyperthermia in pigs has been based on a single QTL (the 'halothane' gene) and the elimination of the responsible RYR^T allele from the

population (Fujii et al. 1991). Further, individual loci have been associated with the genetic variation in resistance to *Escherichia coli* strains in pigs, and have become part of commercial breeding programmes (Jørgensen et al. 2004; The Danish Pig Research Centre Annual Report 2014).

Another consideration for the exploration of the genetic architecture of bTB resistance is that different loci have different properties, i.e. they might act in an additive, or a non-additive way. Assuming a biallelic locus with three possible genotypes (AA, AB and BB), when allele effects are additive, the more copies of the beneficial allele are present in the genotype, the greater the fitness will be (w_{AA} corresponding to the AA genotype in Fig. 1a, where w represents the fitness of the genotype). Through the mechanisms of natural selection, those alleles that confer fitness benefits (i.e. better survival and reproductive success) are expected to proliferate in the population. Therefore, in these cases natural selection will be directional and will move the beneficial allele to fixation (Fig. 2a). When the locus acts in a non-additive way, one of the alleles is dominant over the other, and the heterozygote is more similar phenotypically to the homozygote for the dominant allele (allele A in Fig. 1b). There are different degrees of dominance, and when the heterozygote is identical to the homozygote dominance is complete. In overdominance (or heterozygote advantage) the heterozygote is superior to both homozygotes (w_{AB} in Fig. 1c), and selection is balancing (Fig. 2b). In underdominance (or heterozygote disadvantage) the heterozygote is inferior to both homozygotes (w_{AB} in Fig. 1d) and selection is disruptive (Fig. 2c) (Charlesworth and Charlesworth 2010; Altrock et al. 2011).

Heterozygote advantage implies that a variant is expected to be maintained in the population although it has reduced fitness when homozygote, with the allelic frequency *(q)* reaching a stable equilibrium *(q*)* where Δq (i.e. the change in allelic frequencies per generation) becomes zero. Selection pressure against the alleles for the inferior homozygote is offset by the selection pressure for the allele from the heterozygote. Reversely, heterozygote disadvantage favours extreme values for a trait over intermediate values. The equilibrium allelic frequency has the minimum mean fitness and is unstable. Thus, the population is driven by chance deviations to fixation of one of the two alleles (moving from q^* towards *0* or *1*) depending on their initial frequencies in the population. A classic example of heterozygote advantage is sickle cell anaemia, where although the homozygote genotype for a mutation in the β-globin gene is typically lethal, the heterozygote has better fitness in terms of increased resistance to malaria (Charlesworth and Charlesworth 2010). Another example of heterozygote advantage and balancing selection is the Crooked Tail Syndrome (CTS), where heterozygotes for the causative mutation were showing muscular hypertrophy in a highly selected Belgian Blue cattle population while the homozygote mutants did not survive (Fasquelle et al. 2009). In sheep, the callipyge locus shows a more complex overdominance with the heterozygote individuals (and specifically only the heterozygotes that have inherited the CLPG mutation from their sire, i.e. polar overdominance) expressing muscular hypertrophy (Cockett et al. 1996). Heterozygote disadvantage has been observed for the Rhesus blood group system in humans as well as in the establishment of chromosomal rearrangements (Charlesworth and Charlesworth 2010, Altrock et al. 2011), but has not been encountered before in the context of disease resistance.

The objective of this study was to explore the properties of the individual loci which lead to differences between animals, and to test the hypothesis that it is a single QTL affecting bTB resistance, by means of a Genome Wide Association (GWA) analysis. GWA analysis is a technique that exploits large-scale SNP data to identify associations between genetic polymorphisms and a trait, and it has been widely used in human and animal studies (Andersson et al. 2009). However, non-additive genetic variation is not captured in the standard GWA analysis. Hypothesising underlying non-additive genetic variation, a novel approach is presented to identify loci displaying heterozygote (dis)advantage associated with resistance to bTB and compare such loci with results obtained from standard genome scans.

3.2 Materials and Methods

3.2.1 Data description

The dataset comprises *1,151* Holstein-Friesian cows (*592* confirmed cases and *559* controls with multiple negative tuberculin test results), from *165* herds in Northern Ireland (for a more detailed description of this data see previous chapter). All individuals were genotyped with the 700K BovineHD Illumina Bead Chip and after initial quality control, *617,885* SNPs were retained for subsequent analyses (Table 1).

3.2.2 Constructed datasets

Four different datasets were constructed. Firstly, the full dataset was analysed, comprising all *1,151* cows (Dataset 1).

Secondly, Classical Multidimensional Scaling (CMDS) was conducted, which is a method for visualising the dissimilarities between the individuals based on their genome-wide IBS pairwise distances matrix (*"cmdscale"* and *"as.dist"* functions in "GenABEL", R/2.15.2). CMDS revealed a secondary distinct cluster of *40* individuals (*39* of which originated from the same herd and were possibly crossbreds), as was observed previously for this dataset (see previous chapter section 2.2.4). Therefore, a reduced dataset *(n=1,111)* was constructed after removing the animals clustering separately to address the hypothesis that there might be genetic structure in the data which could potentially lead to false positives (Dataset 2).

Thirdly, a subset of animals *(n= 929)* was derived from the full dataset after removing all the herds that did not contribute any controls (Dataset 3). One herd that contributed only three controls and no cases was retained in the dataset. Lastly, Dataset 3 was further reduced to derive a balanced set *(n= 670)* after randomly removing individuals within each of the remaining herds, so that an equal number of cases and controls would be contributed by each herd (Dataset 4) (Table 1).

3.2.3 Analysis

3.2.3.1 Standard GWA analysis

Standard Genome Wide Association (GWA) analyses were conducted using the GenABEL package (*R version 2.15.2*) (a) for the full dataset (Dataset 1) and (b) for the subset after removing all the herds contributing no controls (Dataset 3).

Firstly, the full population was analysed for SNPs across the genome having an effect on bTB resistance, as what has been shown in a previous study on the same population (Bermingham et al. 2014). After Quality Control (QC) (MAF= *0.05,* callrate = *0.95*), *1,150* individuals and *549,687* markers passed all criteria. Genome wide associations were tested using the *"mmscore"* function in GenABEL taking into account relatedness between individuals as described below. A Q-Q plot was obtained and the genome wide degree of inflation *(λ)* was calculated for the distribution of P-values (*1* d.f. x^2 test statistic) (http://svitsrv25.epfl.ch/R-doc/library/GenABEL/html/estlambda.html). The standard GWA analysis was repeated on Dataset 3 using the same quality control criteria as above, and with the IBS matrix being calculated only for the individuals retained in this dataset.

3.2.3.2 Heterozygote advantage GWA analysis

In order to capture non-additive genetic variation, a modified approach for Genome Wide Association (GWA) analysis was developed with the genotypes being recoded so that there would be only two genotypic classes: heterozygotes (A_1A_2) and homozygotes (including both major and minor allele homozygotes: A_1A_1 and A_2A_2). The same quality control (QC) criteria were applied to all the datasets (MAF= *0.05*, callrate = *0.95*, one female was found to be a male with odds *>1000*, and was discarded). The number of individuals and the number of SNPs that passed QC for each of the constructed datasets and were retained in subsequent analyses can be found in Table 1. The modified GWA analyses were performed on each of the constructed datasets using the GenABEL package (*R version 2.15.2*) as follows:

In analysis 1, the full Dataset 1 was analysed using the "polygenic" and "mmscore" functions in GenABEL. The "polygenic" function was used to estimate the residuals under the polygenic model for the "mmscore" function which takes into account relatedness between individuals. The following model was used:

$$y = \mu + \mathrm{X}\beta + Z\alpha + e \quad (1)$$

where y is the binary bTB status (*0*: control, *1*: case), μ is the overall mean, β is the vector of fixed effects as described in Chapter 1, a is the additive genetic effect derived from the markers with a ~ MVN (0, $\mathbf{K}\sigma_a^2$), and e is the residual error. **K** is the Identity by State IBS kinship matrix of the kinship coefficients representing the probability that an allele sampled with replacement between all pairs of individuals is IBS. **K** was calculated using the "ibs" function in GenABEL, from the markers with the original (i.e. non-recoded) genotypes as follows:

$$f_{ij} = \frac{1}{N}\sum_{k=1}^{n}\frac{(x_{ik} - p_k)(\mathrm{x_{jk}} - \mathrm{p_k})}{\mathrm{p_k}(1-\mathrm{p_k})}$$

where N is the number of SNPs, x_{ik} is a genotype of i^{th} individual at the k^{th} SNP coded as *0*, *1/2*, *1*, and p_k is the frequency of the reference allele (http://www.genabel.org/manuals/GenABEL). Subsequently, the *"mmscore"* function was used to conduct a score test for associations between the SNPs across the genome and the trait under study based on their P-values (http://svitsrv25.epfl.ch/R-doc/library/GenABEL/html/mmscore.html).

In analysis 2, additional approaches were followed to correct for the substructure observed in the CMDS:

(a) Dataset 1 was analysed using the *"egscore"* function which is a score test for association between the SNPs across the genome and the trait under study, adjusting for stratification by principal components derived from the genomic kinship matrix calculated as above (http://www.genabel.org/manuals/GenABEL). Model (1) was followed as described above. Fitting as covariates the two primary axes calculated from the CMDS analysis, was additionally tested. The genome-wide degree of inflation due to hidden substructure *(λ)* was estimated to be *1.000002 (SE= 1.137279*10-5)* (i.e. indicating no serious inflation in the data), nevertheless, it was taken into account for the correction of the distribution of P-values.

(b) Dataset 2 constructed by removing the animals clustering separately, was analysed using the *"polygenic"* and *"mmscore"* functions, and the IBS kinship matrix as described above, calculated only for the individuals retained in the analysis. After removing the minor cluster, λ had a value of *1 (SE= 7.3*10^{-6})* and thus, no further correction for substructure was required.

In analyses 3 and 4, GWA analysis for heterozygote advantage was repeated as described above on Datasets 3 and 4, using the *"polygenic"* and the *"mmscore"* functions, and with the IBS kinship matrix being calculated only for the individuals retained in each of the constructed datasets.

3.2.4 Significance thresholds

Significance thresholds were obtained after the Bonferroni correction for multiple testing as $-log_{10}(0.05/N)$ and $-log_{10}(1/N)$ for the genome-wide and the suggestive thresholds respectively, where N is the total number of SNPs, and as

$-log_{10}(0.05/n)$ for the chromosome-wide significance threshold where n is the number of SNPs on the chromosome. The SNPs of interest were identified as those that exceeded the chromosome-wide and suggestive significance thresholds.

3.2.5 Genotypic frequencies and HWE test

For those SNPs of interest that were identified in the heterozygote advantage analysis, the genotypic frequencies were calculated for each of the genotypic classes and significant loci were tested for Hardy-Weinberg Equilibrium (HWE). Under the assumption of random mating, the HWE test provides information about selection affecting the allele frequencies. Selection might have occurred from conception and before birth, through for example a mutation that is lethal for the homozygotes, or after birth, through reduced fitness throughout life. In those cases the next generation of parents would not be in HWE. However, if mating is random one generation would be enough to return to HWE, thus, departure from HWE informs about selection that has occurred recently. The chi-square test was applied to test for departure of the observed phenotypic frequencies from the HWE expectations, for (a) the full dataset, (b) separately for cases and controls, and (c) pairwise between cases and controls.

3.2.6 Predicted genotypic means

A linear mixed model was fitted in ASReml (Gilmour et al. 2002) with the non-recoded genotypes for the SNP of interest fitted as a fixed effect in order to obtain the predicted mean effects associated with each of the three genotypic classes (A:A = 0, A:G = 1 and G:G = 2). From the analysis were excluded *92* individuals

with missing genotypes for the SNP. Relatedness was accounted for through the IBS relationship matrix (**G**) calculated as in Chapter 1 (see section 2.2.6.2).

Further, Generalised Linear Mixed models (GLMMs) were fitted to take into account the binary (case/control) character of the data, and specifically threshold models (TMs) using the probit and logit link functions (assuming that the underlying liability distribution was normal or logistic respectively), as well as the complementary log-log link function (assuming an underlying extreme value distribution). Additionally, the SNP genotypes were fitted as two covariates aiming to capture additive and non-additive effects for that locus.

Lastly, a linear mixed model with an interaction component was fitted to test for interaction between the SNP identified from the heterozygote advantage analysis and a significant SNP identified by the standard GWAS.

3.2.7 Region exploration and gene expression

As will be described in the results, one SNP in particular was found to be of interest and this SNP was mapped onto the cattle genome using the Ensembl and NCBI genome browsers. The functionality of the gene and its expression pattern were examined through Polymerase Chain Reaction (Roche Taq DNA polymerase) (Jensen K., personal communication, January 18, 2013). Genomic location and sequencing information was obtained using the ENSEMBL browser. Monocytes and macrophages were used to detect expression in three states: (a) resting, (b) at *2* hrs. post-activation with E. coli derived Los, and (c) at *6* hrs. post-activation, and all were compared to negative controls. The PCR consisted of *40* cycles of an initial

dematuration step at *95 ^{o}C*, an annealing step at *40 ^{o}C – 60^{o}C*, and a final extension step at *72^{o}C*.

3.3 Results

3.3.1 Standard GWA analysis

The standard GWA analysis identified a SNP (rs109042660) on chromosome 13 (BTA13) which was significant at the chromosome-wide significance level (*-log$_{10}$(P-value) = 5.57*, while the genome-wide, suggestive and chromosome-wide significance thresholds were *7.06*, *5.76*, and *5.53*, respectively) (Table 2). Additionally, and in agreement with Bermingham et al. (2014), *6* more SNPs on the same chromosome (rs42494342, rs42494357, rs110465273, rs137562332, rs132841890 and rs109809949) were found to be within the ten most significant SNPs, although they did not reach any significance threshold in the present analysis (Fig. 3a, Table 2). However, different quality control criteria and statistical association methods were followed in the two studies. The corresponding Q-Q plot is presented in Figure 3c. GWA analysis on the dataset after removing the herds contributing no controls did not show any significant associations (Fig. 3b).

3.3.2 GWA analysis for heterozygote advantage

GWA analysis for heterozygote advantage on all the animals, identified a SNP on BTA6 (rs43032684) which was significant at the chromosome-wide level (*-log$_{10}$(P-value) = 6.29*) (Table 3, Fig. 4a). The second most significant SNP identified was on BTA25 (rs109960101), but it did not reach significance (*-log$_{10}$(P-value) = 4.94*) (Table 3).

After CMDS analysis, the identified SNP on BTA6 was consistently found to be significant in all the approaches used to control for substructure (Table 3, Fig 4). In the analysis for the reduced dataset after removing the herds contributing no controls, a pattern on BTA6 was still visible and the SNP of interest was the third most significant SNP identified in this analysis but it did not reach suggestive significance ($-log_{10}(P\text{-}value) = 5.25$), while another SNP on BTA17 was now significant at the suggestive level ($-log_{10}(P\text{-}value) = 6.12$) (Table 3, Fig. 5a). GWA analysis for heterozygote advantage on the balanced dataset did not show any significant associations (Fig. 5b).

3.3.3 Genotypic frequencies, HWE test and Genotypic means

Further analysis was done on the single SNP (A/G) of interest (rs43032684) and the genotypic frequencies for the original genotypes (i.e. before recoding for homozygotes and heterozygotes), were calculated for each of the genotypic classes (Table 4). The chi-square statistic was used to test for significant departure from HWE expectations. For the data comprising all the animals, there was a significant departure from HWE ($x^2=10.12, p<0.01$) (Table 5a). When HWE test was performed separately for cases and controls, a significant departure from HWE expectations was observed for the controls ($x^2=28.63, p<0.001$) but not for the cases ($p>0.5$) with the genotypic frequencies in the cases being consistent with the expectations (Table 5b and c). In the controls, the heterozygotes were fewer than expected under HWE (Table 5). The hypothesis that the genotypes would be similarly distributed in cases and controls was tested, to investigate if a case and a control would be equally likely to have a certain genotype with a probability of *0.5*, reflecting the distribution of

cases and controls in the sample. Pairwise chi-square test between cases and controls for each genotypic class, showed that the heterozygotes are significantly fewer than expected under HWE in the controls *(x^2 = 11.50, $p<0.001$)* while the GG homozygotes were more than expected in the controls *($x^2=15.06$, $p<0.001$)*.

The predicted genotypic means corresponding to the linear and the threshold models used are presented in Table 6. The predicted values for the genotypes would be *0* if all the animals in the genotype were healthy and *1* if all the animals were infected. Thus, the values obtained in the present analyses indicate that while the homozygotes show similar values in all the analyses, the heterozygotes show a value more towards *1*, i.e. were more likely to have a diseased phenotype, showing a heterozygote disadvantage pattern which is consistent with results from the HWE test.

The ASReml analysis for rs43032684 provided a Wald F Statistics after adjusting for other effects, of *13.45 ($P <0.001$)*, for the proportion of variation explained by the SNP in the LM model (Table 6). In all the LM or TMs used, the F values show that the SNP genotype had a significant effect on the phenotype of the animal (i.e. being a case or a control). When fitting the SNP genotypes as covariates after adjusting for additive effects, this result was more significant (*$F=17.24$, $P <0.001$*) for the dominance effect of the SNP on the liability scale. The SNP identified was found to explain *1.7%* of the total phenotypic variance. Finally, the interaction with a significant SNP (BTA13) identified by a standard GWA analysis (Bermingham et al. 2014), was not found to be significant *($P >0.1$)*.

3.3.4 Region exploration and PCR

The Ensembl and NCBI genome browsers were used to search the region of the SNP identified through the heterozygote disadvantage GWA analysis, for candidate genes of known function. The SNP (chromosomal position on BTA6: 10,245,091 bp) was found to be adjacent to several candidate genes (Fig. 6). A closer look at the area surrounding the SNP showed that the SNP resides within two partly overlapping Copy Number Variations (CNV) regions (Fig. 7). These CNV regions have been previously associated with resistance to gastrointestinal nematodes in cattle (Fig. 7) (Ensemble, UMD 3.1 assembly; Hou et al. 20011; Hou et al. 2012).

More specifically, the SNP has two alleles (G/A) and resides within the peroxiredoxin-6-like pseudogene (LOC 784039, chromosomal position: 10,223,720-10,246,027 bp, transcript length: *840* bps) (Fig. 8). The pseudogene contains two exons and the SNP resides down-stream of exon 1. Pseudogenes usually derive from a parental gene through the mechanisms of (a) reverse RNA transcription and re-insertion into genomic DNA or (b) through gene duplication. Insertions, deletions or point mutations cause the pseudogenes to be non-functional homologues of the functional parental genes (Gerstein et al. 2006). In order to identify the parental gene of PRDX6L the sequence of the PRDX6L was blasted in Ensemble against the peroxiredoxin-6 gene (PRDX6 on BTA16), which confirmed that PRDX6 is the parental gene of PRDX6L (Jensen K., personal communication).

Initial PCR results showed no clear expression of the pseudogene in monocytes or macrophages, while the parental gene was shown to be expressed. PCR was repeated with monocytes cell templates from five different animals at *50* cycles,

in order to increase amplification of the product. Although the process might have benefited from the use of specific primers, there was an indication of expression of the pseudogene, however, there was no obvious pattern across the different animals (Fig. 9).

3.4 Discussion

3.4.1 Heterozygote disadvantage GWA analysis

Individuals more resistant to bTB can be selected using markers across the genome. In the present study, the hypothesis that effects of individual loci may lead to differences between animals was investigated. Firstly, the standard GWA analysis approach was followed to identify markers associated with bTB resistance, carrying alleles with additive effects. This analysis recovered the same SNPs as presented by Bermingham et al. (2014) for the same population, although the two studies were using different quality control criteria and statistical association methods.

Secondly, in the present study, the possibility of loci with non-additive effects associated with bTB resistance was investigated. Heterozygosity has been linked to better health and improved performance i.e. hybrid vigour, counteracting inbreeding depression which is the loss of heterozygosity (Falconer and Mackay 1997, p. 247; Toro and Maki-Tanila 2007, p. 88; Woolliams 2007, p. 148). Reduced diversity and inbreeding depression have been shown to have negative impacts on fitness traits (Wiener et al. 1994). Conversely, diversity has been suggested to be maintained through heterozygote advantage for example in the case of the Major Histocompatibility Complex (MHC) region, where genetic diversity is beneficial for

the efficiency of the immune system (Codner et al. 2011). These impacts can be observed when there is some degree of dominance between the alleles within a locus, and such an effect would not be detected in standard association analyses when regressing phenotypes on allele counts.

In a previous study (Driscoll et al. 2011) microsatellites were used to test candidate genes for association with bTB susceptibility in particular for effects with heterozygote advantage. However, this study was a smaller study with *384* cattle of multiple breeds including *160* SICCT reactors. Compared to Driscoll et al. (2011), the novelties of this Chapter are: (a) SNPs across the entire genome were used, and (b) the underlying non-additive genetic variation was explored so that both the cases of a heterozygote advantage and disadvantage could be investigated. In the present analysis, there was no evidence for a heterozygote advantage for bTB resistance. Thus, heterozygosity was not found to be beneficial for bTB in cattle. The SNP identified on BTA6 in this study shows a heterozygote disadvantage, suggesting an association between locus heterozygosity and increased susceptibility to bTB. This SNP was significant at the chromosome-wide level after all the methods followed for correction for population structure. Driscoll et al. (2011) identified two microsatellite markers that showed heterozygote advantage (INRA111 with chromosomal position on BTA11: 40,311,694-40,311,817 bp, and BMS2753 on BTA9: 76,800,661-76,800,769 bp), associated with bTB susceptibility. This study found no evidence of such SNPs on chromosomes 11 or 9, and the most significant SNP on either BTA9 and BTA11, was rs110974556 on BTA9 with chromosomal position 40,944,275 bp, which is ~*36* Mbp away from BMS2753, and which did not approach significance.

For further quality control, the SNP genotype call graph for the SNP showing heterozygote disadvantage was examined for any ambiguity in the SNP genotypes, which could have an undue impact on the results. The three genotypic classes for the SNP were found to form three distinct clusters, indicating good genotyping accuracy (Fig. 10).

Significantly fewer than expected heterozygotes for rs43032684 were found in the controls i.e. heterozygotes were more likely to be diseased. The magnitude of the observed difference for the heterozygotes is indicative of dominance effects, and suggests a fitness disadvantage for the heterozygotes. It would be a disadvantage (and not an advantage) because the departure from HWE is observed in the controls i.e. in healthy individuals. The dominance hypothesis was confirmed through the ASReml analysis which provided a significant F-test statistic. The predicted genotypic means confirmed the heterozygote disadvantage for the heterozygotes which showed a value more towards *1*, i.e. more likely to be diseased, which is consistent with the results from the HWE test. The heterozygote advantage GWA analysis presented in this study is not restricted to identify SNPs showing a heterozygote advantage and it can equally identify a disadvantage, since the recoding of the genotypic classes is selected randomly. This result shows that when adjusted for a "heterozygote abnormality" this SNP is significantly associated with bTB resistance.

Compared to results from standard genome scans there was no interaction observed with a significant SNP identified from a standard GWA analysis (Bermingham et al. 2014) on the same population. The heterozygote disadvantage GWAS presented here, allowed identifying a novel putative QTL. The SNP

identified was found to explain *1.7%* of the total phenotypic variance, which although sounds quite small, in comparison to results from standard GWAS studies, this is a relatively large effect. However, this could be an overestimate of the proportion of variance explained by the SNP due to the inherent error in the estimate of the genotypic values. The sample size has a large impact on the power of the GWA analysis to detect associations (Spencer et al. 2009) and further studies are needed to confirm these findings and validate the QTL on larger case-control datasets.

3.4.2 Biological interpretation

The SNP showing heterozygote disadvantage was found to reside within a pseudogene. Understanding the properties of pseudogenes can give insight into how the identified SNP and pseudogene might have a biological meaning. Pseudogenes are non-functional homologs of a functional gene (Vanin 1985). They are defective due to genetic lesions such as insertions, deletions and premature stop codons that do not allow them to encode functional polypeptides. However, they may maintain some functionality and be involved in gene expression regulation (Korneev et al. 1999; Hirotsune et al. 2003; Kandouz et al. 2004). There are non-processed pseudogenes which are mutated duplicates of a parental gene, or processed pseudogenes which are the products of reverse mRNA transcription (Vanin 1985; Bischof et al. 2006). Processed pseudogenes usually occur on a different chromosome compared to the functional parental gene (Vanin 1985). The pseudogene identified in the present study on BTA6 was found to have a parental gene (PRDX6) on BTA16 and thus, more likely to be a processed pseudogene. The product of the parental gene PRDX6

is an enzyme (peroxiredoxin-6) involved in lipid internalising and degradation. It is expressed in the respiratory epithelium in the lungs, in alveolar type II cells and in alveolar macrophages, and is localised in the lysosomes (Sorokina et al. 2009; Chatterjee et al. 2011). A peptide within the PRDX6 protein sequence (amino acids 31-40), necessary for the lysosomal localisation of the protein, was found to be *100%* conserved in humans, rats, mice and cattle (Sorokina et al. 2009).

Mycobacteria, in order to overcome the immunological reaction of the host and survive within the macrophages, manipulate host signalling processes through mediator molecules. These mediator molecules are lipids and glycolipids released from the bacterial cell-wall that accumulate within the host's macrophages and into the membrane of the phagosome containing the internalised mycobacteria (the process called phagocytosis) (Fig. 11). Mycobacterial mediator lipids interfere with the fusion of the phagosome with the lysosome, which contains enzymes and would result in killing the mycobacteria, therefore inhibiting the mycobacterial killing processes (Anes et al. 2003; Koul et al. 2004; Raman et al 2010). The capacity of phagocytosis and the rate of intracellular killing of mycobacteria have a strong influence on the outcome of infection i.e. controlled or uncontrolled infection (Gammack et al. 2004), and blocking of phagocytosis and phagosomal maturation has been shown to impair clearance of infection (Raman et al. 2009). In the PCR there was clear expression of the parental gene in monocytes. For the pseudogene there was indication of expression but with no clear pattern across the different animals, and thus PCR results were not conclusive. Therefore, although in this study expression of the pseudogene was not confirmed, given the role of its parental gene in lipid internalising, hypothetical functionality of the pseudogene could have an

effect on the internalising of the virulence factors, inhibiting mycobactericidal reaction.

3.4.3 Conclusion

A SNP was identified on BTA6 suggesting an association between locus heterozygosity and increased susceptibility to bTB in cattle, and implying a fitness disadvantage for the heterozygotes at this locus. The SNP resides within a CNV region associated to nematode resistance in cattle. Further, the SNP was found to reside within a pseudogene. However, despite the apparent functional relevance, it was not possible to demonstrate functionality of the pseudogene. The novel method presented in this Chapter for identifying loci displaying heterozygote (dis)advantage associated with disease resistance, captures non-additive genetic variation and allows identification of associations that in a standard GWA analysis would go undetected.

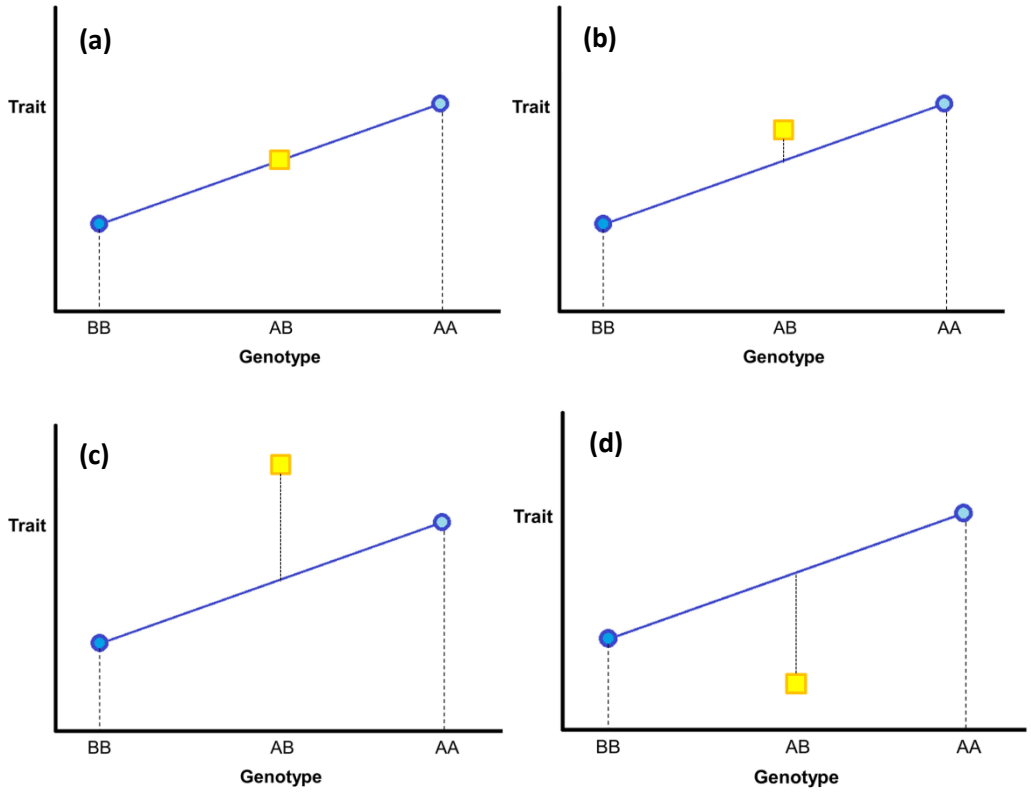


Figure 1. Graphs representing the different QTL properties: (a) additive model *((AA – BB) / 2)*, (b) dominance *(AB - (AA + BB) / 2)*, (c) overdominant heterozygote advantage and (d) underdominant heterozygote disadvantage.

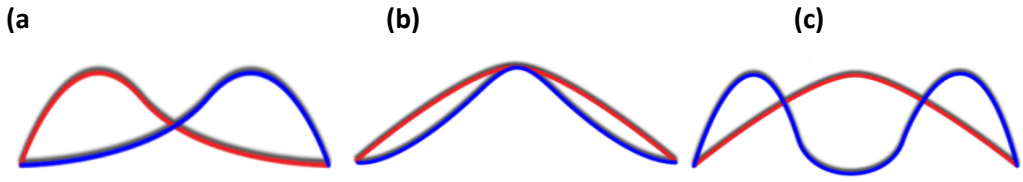


Figure 2. The effects of the three types of selection on the genotypic frequencies before (red line) and after (blue line) selection: (a) directional selection, (b) balancing selection, and (c) disruptive selection (http://en.wikipedia.org). On the X axis are represented the genotypic values and the Y axis represents the relative frequencies.

Dataset	Animals	Cases	Controls	SNPs	Animals after QC	SNPs after QC
1	1151	592	559	617885	1150	549687
2	1111	552	559	617885	1110	549835
3	929	370	559	617885	929	550108
4	670	335	335	617885	669	550502

Table 1. The number of animals and number of SNPs retained in the analyses after Quality Control (QC) for each of the constructed datasets: (1) Dataset 1 is the full dataset, (2) Dataset 2 is reduced by removing the animals clustering separately in the CMDS, (3) Datset 3 is reduced by removing the herds with no controls, and (4) Dataset 4 is balanced with an equal number of cases and controls in each herd.

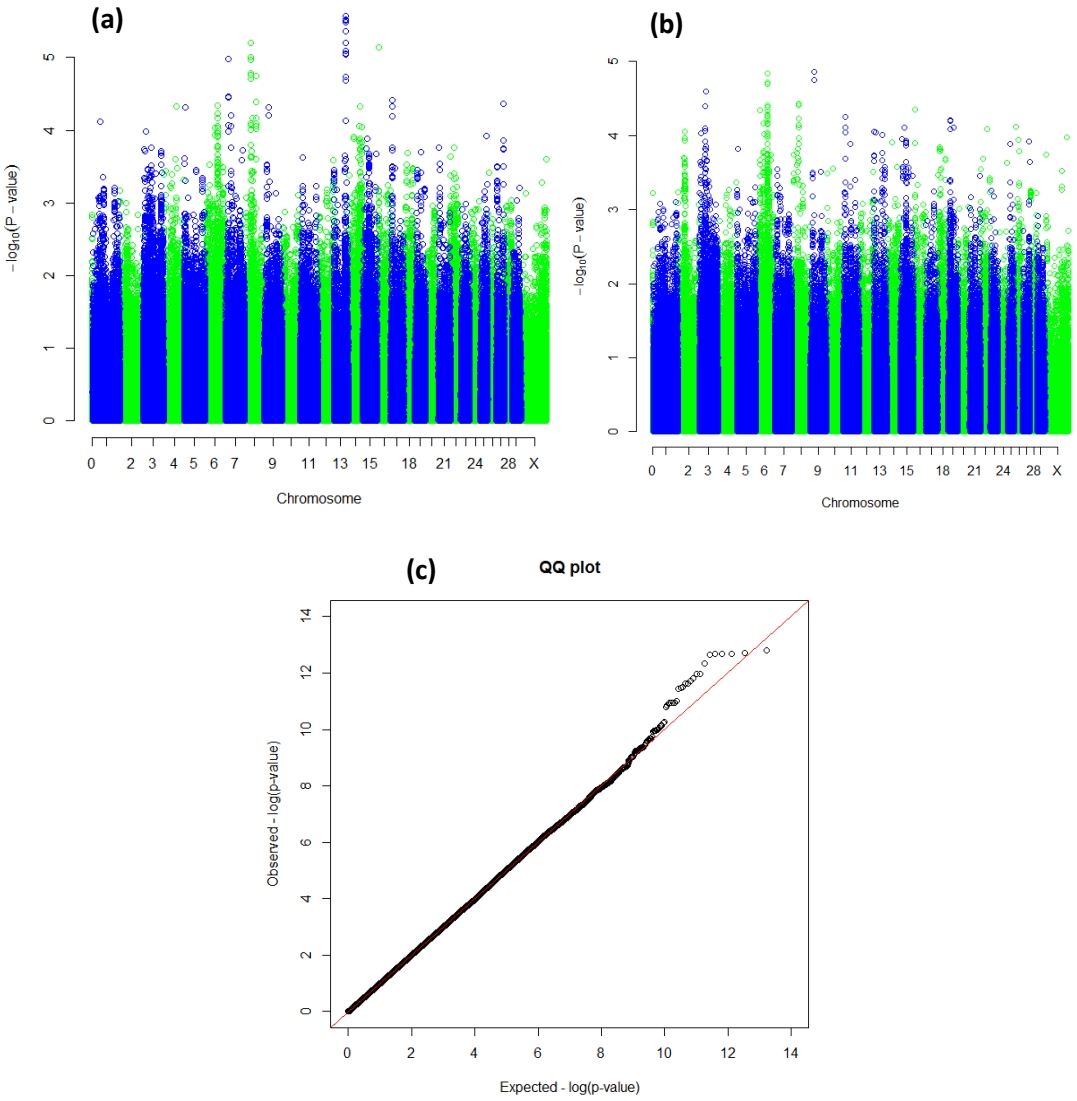


Figure 3. Manhattan plots from the standard GWA analysis on (a) the full dataset and (b) after removing herds contributing no controls; (c) Q-Q plot showing observed compared to expected x^2 values under the null hypothesis of no association.

SNP		Chr	- log_{10} (P-value)
BovineHD1300020589	rs109042660	13	**5.57**
BovineHD1300020586	rs42494342	13	5.52
BovineHD1300020584	rs42494357	13	5.50
BovineHD1300020585	rs110465273	13	5.50
BovineHD1300020590	rs137562332	13	5.50
BovineHD1300020591	rs132841890	13	5.49
BovineHD4100010384	rs43705552	13	5.36
BovineHD1300020588	rs109809949	13	5.20
Genome-wide threshold			**7.06**
Suggestive threshold			**5.76**
Chromosome-wide threshold			**5.53**

Table 2. The *8* most significant SNPs identified from the standard GWA analysis on Dataset 1 comprising all the animals, and corresponding P-values.

Analysis	rs43032684 BTA6	rs109960101 BTA25	rs109682541 BTA17
1	6.29	4.94	-
2a	6.43	-	-
2b	6.09	5.05	-
3	5.25	-	6.12
Genome-wide threshold	**7.09**	**7.09**	**7.09**
Suggestive threshold	**5.79**	**5.79**	**5.79**
Chromosome-wide threshold	**5.78**	**5.34**	**5.58**

Table 3. Association P-values for the SNPs of interest from the GWA analyses for heterozygote advantage on (1) Dataset 1 comprising all animals, (2a) Dataset 1 taking structure into account by fitting PCs, (2b) Dataset 2 formed by removing the animals in the minor cluster from the CMDS, and (3) Dataset 3 formed by removing the herds contributing no controls. The genome-wide, suggestive and corrsponding chromosome-wide significance thresholds are obtained after the Bonferroni correction for multiple testing.

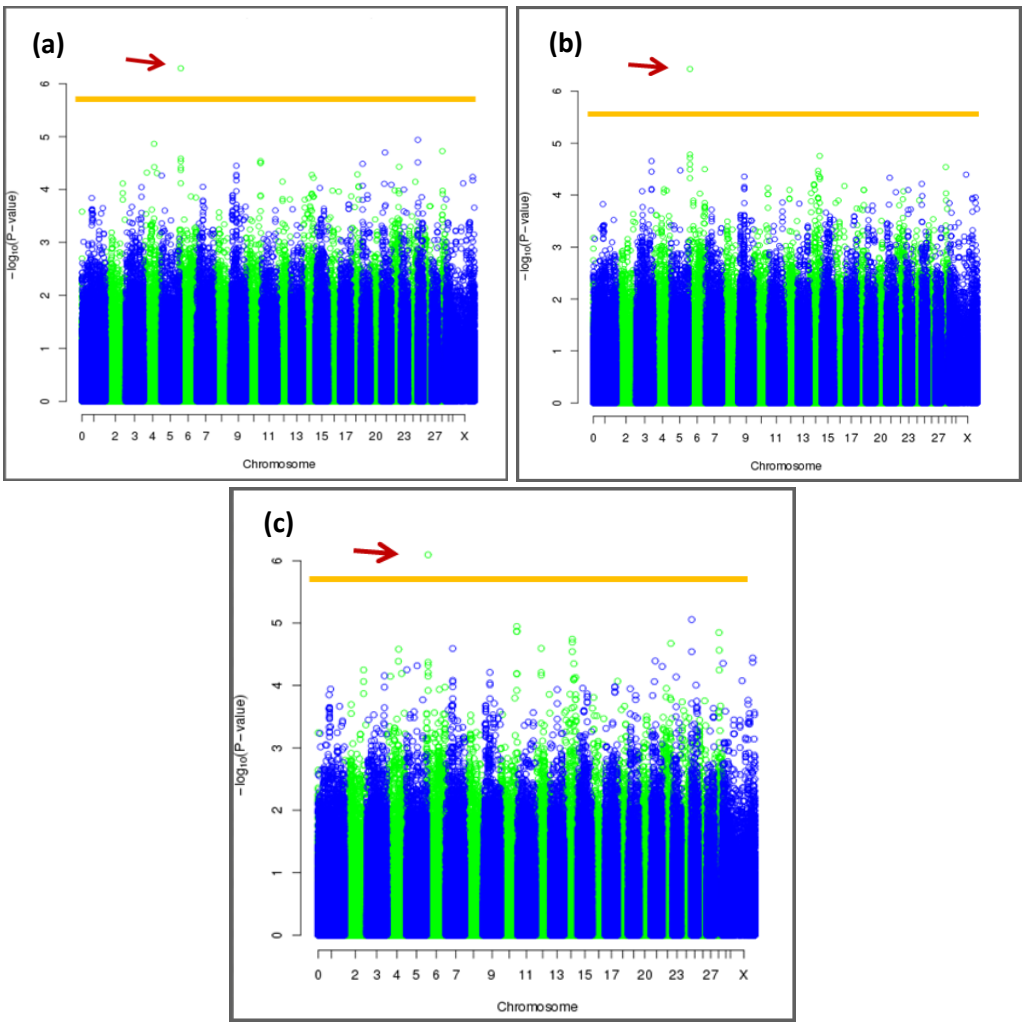


Figure 4. Manhattan plots of the SNP associations with the trait of interest, from the GWA analyses for heterozygote advantage on (a) Dataset 1 comprising all animals (Analysis 1), (b) Dataset 1 taking structure into account by fitting PCs (Analysis 2a), (c) Dataset 2 after removing the animals in the minor cluster from the CMDS (Analysis 2b). The yellow line represents the suggestive significance threshold *(5.79)*.

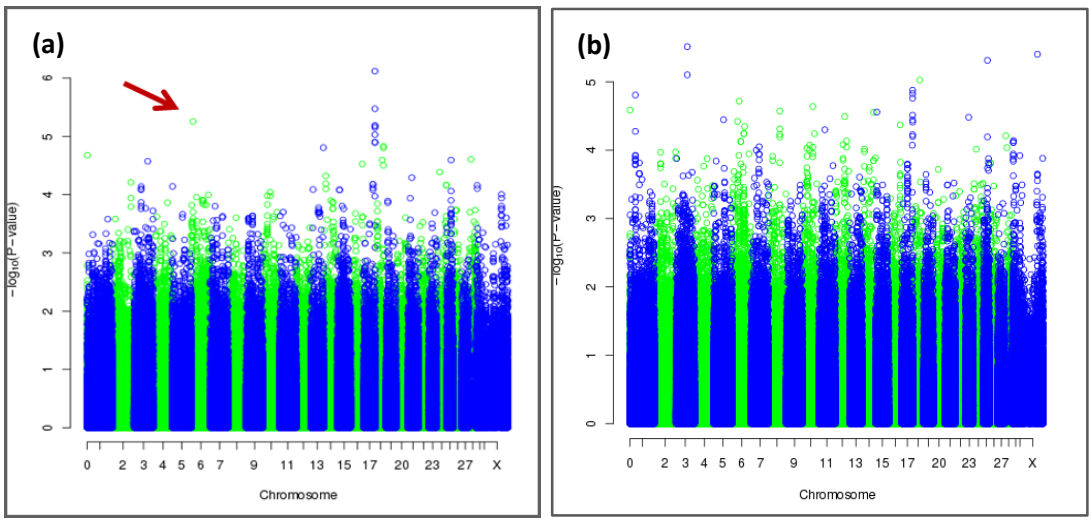


Figure 5. Manhattan plots of the SNP associations with the trait of interest, from the GWA analyses for heterozygote advantage on (a) Dataset 3 after removing the herds contributing no controls , and (b) Dataset 4 with an equal number of cases and controls in each herd.

rs43032684	Cases	Controls	Total	Fraction of total
A/A	239	286	525	0.46
A/G	254	154	408	0.36
G/G	62	63	125	0.11
Total	555	503	1058	-
NA	36	56	92	0.08

Table 4. Genotypic frequencies for the SNP showing heterozygote disadvantage identified in Table 3, for the cases and the controls.

(a)	AA	AG	GG	Total	P-value	p_A
Observed	525	409	125	1059	<0.01	0.69
Expected	502.73	452.38	102.43		x^2=10.12	

(b)	AA	AG	GG	Total	P-value	p_A
Observed	239	255	62	556	>0.7	0.66
Expected	241.46	249.16	64.27		x^2=0.24	

(c)	AA	AG	GG	Total	P-value	p_A
Observed	286	154	63	503	<0.01	0.72
Expected	261.56	201.904	38.87		x^2=28.63	

Table 5. Observed and expected under HWE genotypic frequencies for the SNP showing heterozygote disadvantage for (a) all the animals, (b) only the cases, and (c) only the controls.

	LM (SE)	Probit (SE)	Logit (SE)	Comp. log-log (SE)
GG	0.59 (0.05)	0.61 (0.05)	0.60 (0.05)	0.62 (0.05)
AG	0.72 (0.03)	0.73 (0.03)	0.73 (0.03)	0.76 (0.03)
AA	0.55 (0.03)	0.56 (0.03)	0.56 (0.03)	0.56 (0.03)
F	13.45 (P < 0.001)	13.74 (P < 0.001)	12.64 (P < 0.001)	15.22 (P < 0.001)

Table 6. Predicted genotypic means for the three genotypic classes for the SNP of interest obtained from the ASReml analyses using a linear model and threshold models (values transformed back to the observed scale), with the corresponding Wald F statistic obtained after adjusting for other effects.

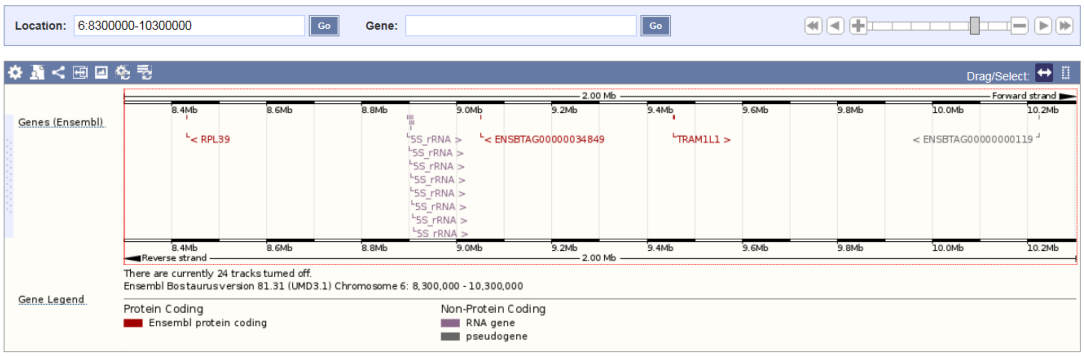


Figure 6. Ensemble genome browser region 6:8,300,000-10,300,000 bp containing the SNP showing heterozygote disadvantage (SNP position: 10,245,091 bp).

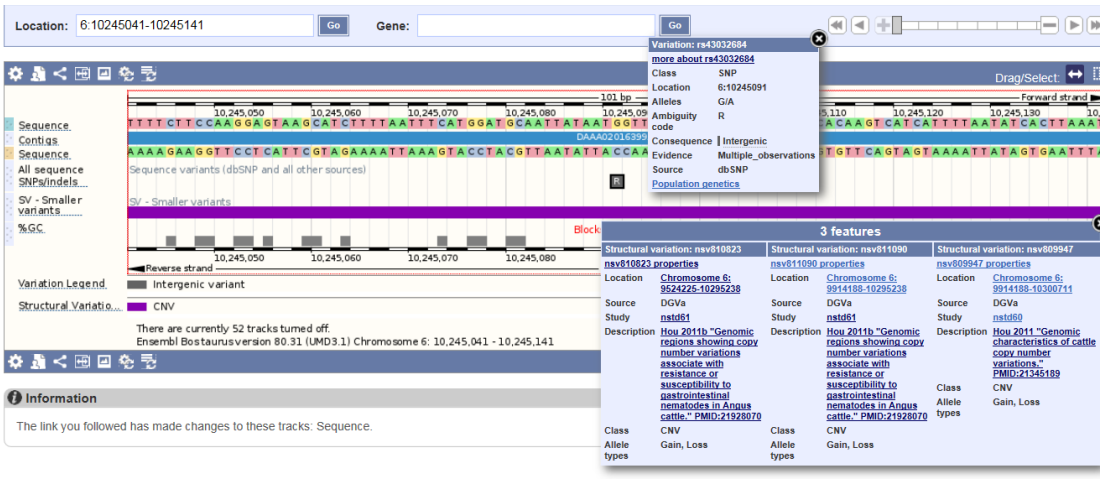


Figure 7. Ensemble genome browser, region: 6: 10,245,041-10,245,141 bp containing the SNP of interest as well as three previously identified Copy Number Variation (CNV) regions associated with gastrointestinal nematodes resistance in cattle.

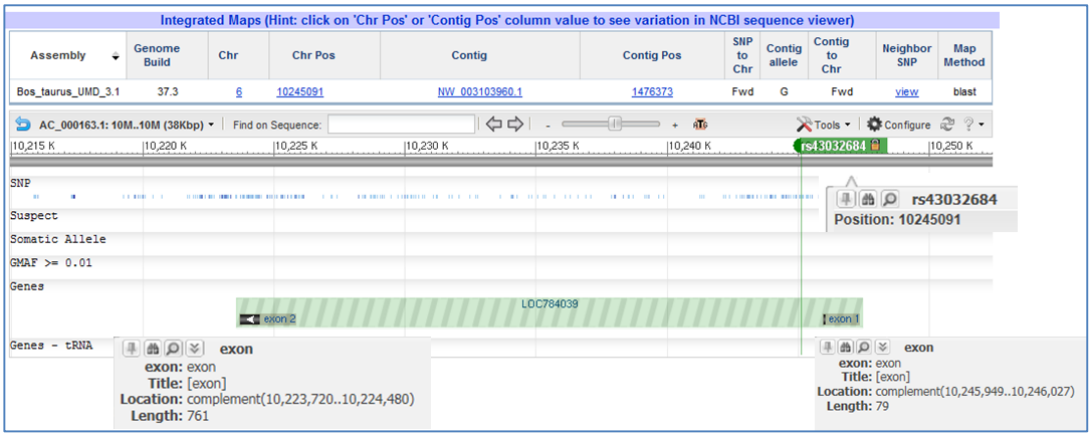


Figure 8. Location of the peroxiredoxin-6-like pseudogene (LOC784039) containing two exons, and position of the SNP identified in the heterozygote disadvantage GWA analysis (rs43032684) residing down-stream of exon 1 (http://www.ensembl.org/index.html).

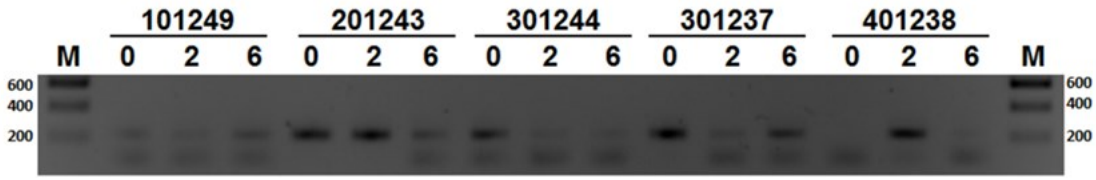


Figure 9. RT-PCR amplification of PRDX6L from bovine monocyte cDNA, for six different animals (0, 2 and 6 denote time post activation in hours with E. coli-derived LPS).

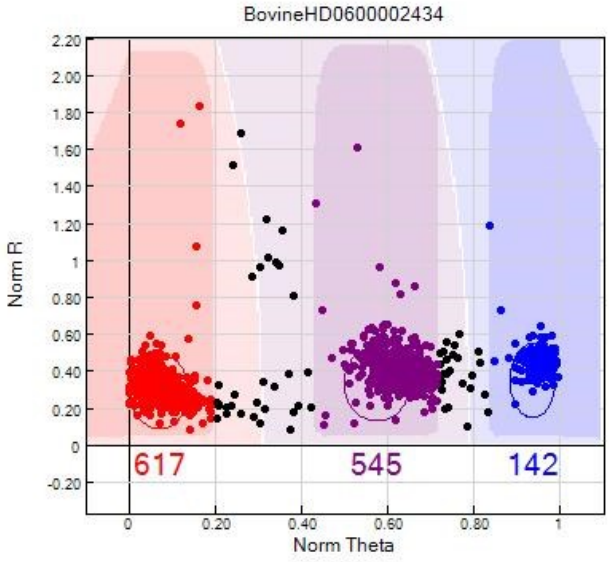


Figure 10. Genotype calls scoring graph for sequencing quality control for the SNP showing heterozygote disadvantage, where the samples are displayed in three distinct shaded areas based on their genotype calls. The three genotypic classes are represented by three distinct clusters (samples in the red region are AA genotypes, samples in the purple region are AG, and samples in the blue region are GG). Black dots represent genotypes that could be assigned to any of the three classes.

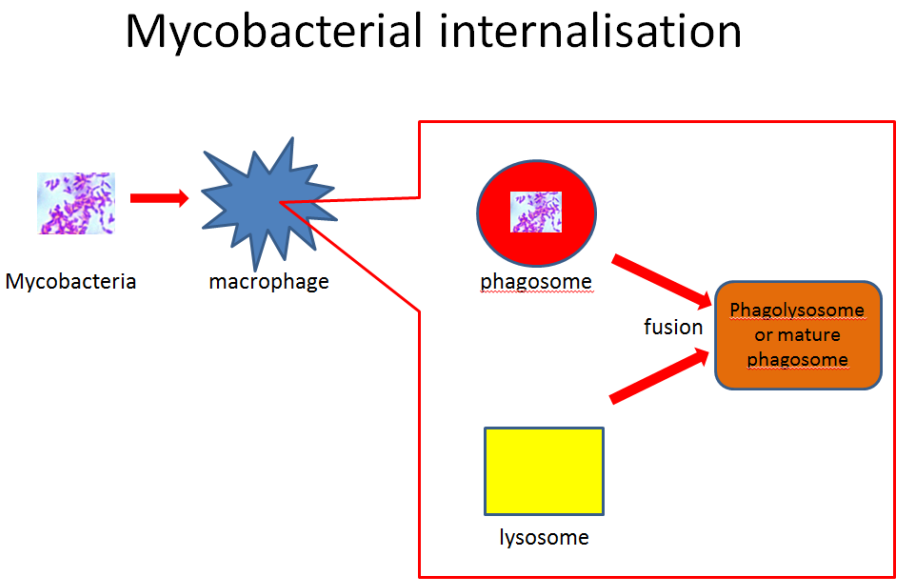


Figure 11. Graphical representation fo the process of infection with *M. bovis*. Macrophages are the primary target cells of mycobacteria. When mycobacteria are internalised within the macrophages in a structure called phagosome, effective immune response and deactivation of the pathogen requires fusion of the phagosome with the lysosomes which are sub-cellular structures containing enzymes. After successful fusion, the phagolysosome (or mature phagosome) is formed, where the pathogen can be destroyed.

Chapter 4

A meta-analysis for bovine tuberculosis resistance in dairy cattle

4.1 Introduction

Genetic selection of individuals resistant to bovine Tuberculosis (bTB) could offer a complementary strategy for the control of tuberculosis in cattle. Previous studies have shown the presence of host genetic variation underlying resistance to bTB (Brotherstone et al. 2010; Bermingham et al. 2012; Tsairidou et al. 2014) and several loci have been associated with individual variation in bTB resistance (Bermingham et al. 2014; Finlay et al. 2012). Selection could either be directly on animal phenotypes (traditional phenotypic selection using pedigrees), or on genetic markers. There are two approaches for using genetic markers, either selection on a small number of loci, or genomic prediction using a large number of markers spread throughout the genome. However, with resistance to bTB being a complex trait and therefore likely to be influenced by a large number of genes, further studies are needed in order to identify additional loci influencing the trait and make MAS more efficient. In chapter 1 of this Thesis, it was shown that genomic selection is feasible; breeding values were estimated and their prediction accuracy was tested in the absence of disease phenotypes (Tsairidou et al. 2014). However, before implementation, the prediction accuracy needs to be improved. Therefore, more information is required both for testing loci and for improving the prediction accuracy.

The identification of variants underlying the control of bTB resistance and the accuracy of genomic prediction are both likely to benefit from larger sample sizes. This may be achieved through meta-analyses which combine results from different studies using potentially different populations. By means of simultaneous analysis of individuals distantly related, assisted by a larger sample size, this approach may reveal new information concerning the genetic architecture of bTB resistance. This approach, i.e. the simultaneous analysis of distantly related individuals, has been found to be powerful in both animal and human studies (Sanna et al. 2008; Willer et al. 2008; Riggio et al. 2014), where it has been used to confirm the presence of QTLs previously identified and to detect new loci affecting the trait under study.

The aim of this study was to investigate the genomic control of bTB resistance and to explore the feasibility of genomic selection for bTB resistance, combining information across datasets in a meta-analysis. Specifically, we anticipate that meta-analyses will provide additional information on specific loci affecting resistance and it will enable enhanced genomic predictions of resistance. This was done by combining data from two independent populations and analysing this joint dataset using different approaches, ranging from Regional Heritability (RH) mapping, through chromosomal heritability estimation, to whole genome prediction both within and across populations.

4.2 Materials and Methods

4.2.1 Description of data

4.2.1.1 Animals and phenotypes

Two populations were used in the analyses: Population 1 comprised *1,151* female Holstein-Friesians originating from commercial herds in Northern Ireland, confirmed bTB cases (positive to the SICCT and confirmed by post mortem examination) or controls (animals that provided multiple negative test results) (Bermingham et al. 2014); Population 2 comprised *287* Holstein and Friesian bulls from the Republic of Ireland with estimated breeding values (EBVs) calculated from their daughter phenotypes for SICTT (Finlay et al. 2012). Each bull had two EBVs, with and without pedigree information used in their estimation, each with a reliability, and a set of SNP genotypes as will be described below. In Population 2, the number of daughters per bull varied, ranging from *4* to *1046* daughters, with a mean of *35* daughters per bull (Fig. 3) and reliabilities without pedigree ranging from *0.03* to *0.9*.

4.2.1.2 Phenotypes

To combine the two populations and analyse them together, an initial fixed effects model was used for the first population, pre-correcting for all known non-genetic fixed effects (age, year of testing, season, reason for testing, and breed) and the residuals of this model were used as phenotypes in subsequent analyses. For Population 2, the de-regressed EBVs (i.e. EBVs divided by their reliability) were used as phenotypes (Fig. 1), either with pedigree information included in their

estimation (pedigree-derived de-regressed EBVs) or without (pedigree-free de-regressed EBVs). All phenotypes were standardised by their origin-specific standard deviation in order to be analysed simultaneously.

4.2.1.3 Genotypes

Population 1 was genotyped using the Illumina high-density Bead Chip, while Population 2 was genotyped with the Illumina Bovine50 SNP chip. To combine genotypes from these two populations, the Illumina forward strand genotypes were obtained for both populations and used to construct the pooled datasets. In total, genotype data from *777,962* SNPs were available for the *1,151* cows of Population 1 and *54,001* SNPs were available for the *287* bulls of Population 2 (Table 1).

4.2.1.4 Population structure exploration

Principal component analysis was conducted in *R* (*R version 2.15.2*), using the IBS matrix calculated on all the *1,438* individuals and using only the SNPs present in both the high density and the low density SNP chips. Plotting the first versus the third principal component, captured the structure previously observed when analysing the two populations independently (Bermingham et al. 2014 Fig. S1; Finlay et al. 2012 Fig. 1), while plotting the first versus the second principal component showed no structure (Fig. 2). The bulls in Population 2 were found to follow a similar pattern to Population 1 and no population-specific clustering was observed, however there was a secondary minor cluster of animals, possibly crossbreds (see Chapter 2, section 2.2.5).

4.2.1.5 Constructed datasets

Two datasets were constructed for subsequent analyses. Dataset 1 comprised all animals (*n=1,438*), where the SNPs present in both SNP chips were used (i.e. *36,690* autosomal SNPs). These SNPs were retained after conducting quality control on Population 1 (MAF<*0.05*, call rate<*95%*, HWE p<*0.000001*, all SNPs homozygote, all SNPs heterozygote, or all missing were removed) and accepting all genotypes for Population 2 (QC1, Table 1).

Dataset 2 was constructed after removing the bulls with less than *8* daughters. After deregressing the EBVs, bulls with very few daughters might have an undue influence on the data, possibly adding noise rather than information. In breeding value prediction from progeny records, reliability is $n/(n+k)$ where n is the number of offspring and k can be calculated as $k = (4 - h^2) / h^2$ (Mrode 2005, p. 7). If the true heritability of bTB resistance is *~0.2*, as indicated by several field studies (Brotherstone et al. 2010; Bermingham et al. 2012; Bermingham et al. 2014; Tsairidou et al. 2014), then to achieve a reliability of *>0.3* at least *8* daughters would be needed. Therefore, the bulls with *<8* daughters which had reliabilities ranging from *0.03* to *0.08,* were removed, and *175* bulls were retained in subsequent analysis in order to address the hypothesis that these individuals might be adding noise. A second round of quality control was carried out (QC2) which applied the same criteria used in Population 1 above, to both Population 1 and Population 2. The final dataset for QC2 comprised *1,326* animals and *34,987* autosomal SNPs (Table 1).

4.2.2 Data analysis

Data analyses comprised investigations of bTB resistance at different levels of the genome, using the combined datasets. These were (i) identification of individual loci influencing resistance, both as individual SNPs and collectively as groups of SNPs, (ii) exploration of the contribution of individual chromosomes to variation in bTB resistance and (iii) the genetic control of bTB resistance at the whole genome level by means of genomic prediction of bTB resistance within and across populations.

4.2.2.1 Regional Heritability (RH) mapping

Regional Heritability (RH) mapping (Nagamine 2012) is a flexible variance component-based means of identifying genomic regions affecting complex traits, particularly when individual SNPs contribute only a small proportion of genetic variation, but groups of SNPs may collectively be significantly associated with the trait. RH can also be an effective means of combining disparate datasets (Riggio et al. 2014a), avoiding the need to assume the same linkage phase between markers and causative mutations across populations. RH was used to identify regions including groups of SNPs collectively affecting bTB resistance. Populations 1 and 2 were initially analysed independently to obtain genomic heritability estimates. For Population 2, pedigree-free or pedigree-derived, de-regressed EBVs were used as phenotypes. An overall genomic heritability estimate for the combined dataset was obtained in an ASReml analysis.

RH mapping methodology was applied as described by Nagamine et al. (2012). The population of origin was fitted as a fixed effect, whereas additive genomic (whole genome) and additive regional effects were fitted as random, with genomic and local IBS relationship (**G, G**$_L$) matrices (Leutenegger et al. 2003) calculated for each window describing the variance/covariance of these effects i.e. $y_i = m + \beta i + u_i + r_i + e_i$, where y_i is the adjusted phenotype of individual i, β_i indicates the population to which i belongs to, u_i its additive genomic effects (**u** ~ MVN(0, σ_u^2**G**)), and r_i its additive regional effects (**r** ~ MVN(0, σ_r^2**G**$_L$)). The data was analysed with three different window sizes: a 50-SNP window size (25-SNP step), 30-SNP (15-SNP step), and 20-SNP (10-SNP step). The RH was calculated for every window as $h^2_r = \sigma^2_r/\sigma^2_p$, where σ^2_r was the variance explained by the window and σ^2_p was the phenotypic variance. Putative regions associated with bTB resistance were identified through a Likelihood Ratio Test (LRT), tested for every window against the null hypothesis of only a polygenic inheritance described by the genomic matrix **G**. The regions with the maximum RH estimate and LRT (RH_{max} and LRT_{max}) were identified for each chromosome. Suggestive and genome-wide significance thresholds were obtained after the Bonferroni correction for multiple testing.

Analyses were repeated for both pedigree-free and pedigree-derived EBVs for Population 2 and results were compared. For completeness, RH mapping was repeated for Dataset 1 using the overlapping SNPs between the HD and the low density SNP chips but after applying QC2 (section 4.2.1.5). This more stringent quality control resulted in the removal of additional SNPs so that *35,021* autosomal SNPs were retained in this analysis (Table 1). In this analysis, the pedigree-free de-regressed EBVs were used as phenotypes for Population 2.

Dataset 2, i.e. where bulls from Population 2 with less than *8* daughters were removed, was analysed in the same way as Dataset 1 with the difference that Population 2 pedigree-derived EBVs were not considered.

4.2.2.2 Genome-wide Association (GWA) analyses

GWA analysis was performed to identify individual loci associated with bTB resistance in the combined populations. A mixed linear animal model was fitted in GenABEL (*R version 2.15.2*) for Dataset 1 comprising all *1,438* individuals, and *35,459* SNPs. The *"check.marker"* function in GenABEL was used for further quality control and SNPs with a minor allele frequency of *<0.05*, with *>5%* missing data, or significantly out of Hardy–Weinberg Equilibrium (*fdrate=0.2*) were excluded from subsequent analysis. Samples with *>5%* missing SNPs or with *>95%* IBS were also discarded. In total *1,438* individuals and *35,286* SNPs passed all QC criteria and were retained in the analysis. The population of origin and three principal components were fitted as fixed effects to account for population structure and so that the combination of principal components used would account for the outlier group of animals (potential crossbreds) previously observed in Population 1. The model was fitted as follows:

$$\boldsymbol{y} = m\mathbf{1} + \boldsymbol{p} + PC_1 + PC_2 + PC_3 + \boldsymbol{u} + \boldsymbol{e}$$

where p_i indicates the population of origin, PC_1, PC_2 and PC_3 are the three principal components and u_i is the genomic estimated breeding value with $\mathbf{u} \sim \text{MVN}(0, \mathbf{G}\sigma_a^2)$. The P-values were those obtained after accounting for the genome-wide

degree of inflation (λ), but adjustments were negligible as the value of λ was *1* (*SE*= $8.95x10^{-5}$) indicating a successful correction for substructure through fitting principal components. The genome wide and suggestive significance thresholds were obtained after the Bonferroni correction for multiple testing as $-log_{10}(0.05/N)$, and $-log_{10}(1/N)$ respectively, where *N* was the total number of SNPs. The phenotypes used for analysis were the pre-corrected phenotypes for Population 1, and the pedigree-free deregressed EBVs for Population 2, standardised by their population specific standard deviation.

4.2.2.3 Chromosomal heritability estimation

Chromosome-level heritabilities were calculated to help interpret the genomic predictions, as this approach gives insight into properties of the genomic heritability and the proportion of genetic variance due to population structure (Yang et al. 2011). Chromosomal heritability estimation follows the same procedure as RH mapping, except that variation is partitioned at the level of the whole chromosome, rather than the SNP window. For the combined Datasets 1 and 2 the heritability for each chromosome was calculated as described below:

Method (a). Heritabilities were calculated separately for each chromosome ($h_{c(sep)}^2$) using single-chromosome specific $\boldsymbol{G}_c$ matrices and fitting the following model in an ASReml analyses: $y_i = \beta + \boldsymbol{g_c} + e$, where β indicates the population of origin and $\boldsymbol{g_c}$ represents the genetic effects of chromosome *j* with $\mathbf{g_c} \sim$ MVN(0, $\sigma_c^2\mathbf{G_c}$), where *c* goes from *1* to *29*. This was done for both Datasets 1 and 2.

Method (b). All variance components (29 chromosomes) were fitted as random effects simultaneously, using an in-house built restricted maximum

likelihood (AIREML) programme, with starting values obtained from the variance estimates in (a). The heritability (h_c^2) was estimated as the variance corresponding to each chromosome divided by the phenotypic variance after fitting the model: $y = \beta + \sum \boldsymbol{g}_c + e$ with $\mathbf{g}_c \sim \text{MVN}(0, \sigma_c^2 \mathbf{G}_c)$. This was done for Datasets 1 and 2.

Method (c). For Dataset 2 each chromosome was fitted separately with distributional assumption for $\boldsymbol{g}_c$ as in method (a), plus a polygenic term representing the remaining *28* chromosomes, (i.e. genomic-chr$_i$), $y = \beta + \boldsymbol{g}_c + \boldsymbol{g}_{-c} + e$, with $\mathbf{g}_{-c} \sim \text{MVN}(0, \sigma_c^2 \mathbf{G}_{-c})$ where $\boldsymbol{G}_{-c}$ was calculated on the SNPs complementary to $\boldsymbol{g}_c$. For each chromosome, the marginal contribution of genetic variance explained by each chromosome was tested (H_0). The Genomic matrix excluding the chromosome under study was used, i.e. $y = \beta + \boldsymbol{g}_{-c} + e$, and LRT was conducted (where L_1 was the log-likelihood from the full model for each chromosome and L_0 was the log-likelihood from the reduced model) to test for significant improvement when the chromosome was included in the model. A similar approach was followed for Dataset 1 however each chromosome was fitted plus the whole **G** matrix. Excluding each time the chromosome under study from the **G** matrix (i.e. approach followed for Dataset 2) is considered to be the optimal analysis.

Method (d). For Dataset 2 only the chromosomes with non-zero variance in Method (c) were fitted, simultaneously, using the AIREML programme with starting values from Method (a).

Method (e). For Dataset 2 only the chromosomes with non-zero variance when all the *29* chromosomes were analysed simultaneously in Method (b), were fitted together with the $\mathbf{G}_{-c}$ matrix on the remaining chromosomes. The proportion of variance explained by the selected chromosomes was calculated as the sum of the

variance explained by these chromosomes over the total phenotypic variance in the same analysis.

Method (f). For Dataset 1, the chromosomes with non-zero variance in Methods (a), (b) and (c) were fitted simultaneously.

Under the assumption of the infinitesimal model for the trait of interest and genetic effects distributed randomly across the genome when fitted separately, the magnitude of the chromosomal heritability should be proportional to the length of the chromosome. However, with genetic structure (relatedness) within the population, the estimated genetic effects will not be independent on different chromosomes. Therefore, the chromosomal heritability estimation analysis allows us to explore the origin of the accuracy of genomic prediction, i.e. if this accuracy is due to markers tagging closely true QTLs or if a proportion of the accuracy is due to relatedness, whereby markers on one chromosome may capture effects on other chromosomes. The proportion of genetic variation that can be attributed to population structure (relatedness) can be inferred from the regressions of chromosomal heritability on chromosomal length (Yang et al. 2011). For both Datasets 1 and 2, $h_{c(sep)}^2$ and $h_{c(sep)}^2$-h_c^2 were regressed on chromosome length (L_c), and the proportion of genetic variance due to population structure was then calculated as $b_0/b_{0(sep)}$, where b_0 and $b_{0(sep)}$ are the intercepts of the two regressions (Daetwyler et al. 2012).

4.2.2.4 Genomic prediction

The accuracy of the within-population and across-populations genomic prediction was tested by means of Cross Validation (CV) (Luan et al. 2009). Two approaches were followed: (a) individuals from both populations were combined and

then randomly assigned to five groups of near-equal size irrespective of their population of origin (5-fold CV), each time using one group as the validation set and the remaining four groups as training sets, and using the model $\boldsymbol{y} = \boldsymbol{m1} + \boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{u}$ is the vector of genomic estimated breeding values with $\mathbf{u} \sim \text{MVN}(0, \sigma_u^2\mathbf{G})$. This procedure was replicated *50* times, each time with a different randomisation of the individuals in the groups. (b) Prediction accuracy was estimated across populations using either Population 2 or Population 1 as the validation set.

The average accuracy across *50* randomisations was calculated in (a), and the expected accuracy in (b), as $E[r(g,\hat{g})] \approx r(y,\hat{y})/h$, where $r(y,\hat{y})$ is the correlation between the cross-validated predicted EBVs ($\hat{y}$) and the phenotypes and h is the square root of the corresponding heritability; the heritability used was within fold for (a) and within population for (b). The standard error of the accuracies in (a) was calculated as the empirical standard deviation of the *50* accuracy estimates.

These methods were applied to both Datasets 1 and 2. Further, in Dataset 1, the procedure was repeated using pedigree-derived EBVs for Population 2, and the results were compared with those obtained when pedigree information was not taken into account in the calculation of the EBVs.

4.3 Results

4.3.1 Genomic heritability estimates

Genomic heritability estimates, for the different datasets and trait definitions are presented in Table 2. Population 1 yielded a heritability on the observed scale of *0.23 (SE = ±0.06)*. For Population 2 heritability was zero when the pedigree-free de-

regressed EBVs were used as phenotypes, and *0.11 (SE = 0.1)* when the pedigree derived de-regressed EBVs were used. When retaining only the *175* bulls with ≥*8* daughters for Population 2, the genomic heritability was *0.60 (SE = 0.22)*, however, the smaller sample size contributed to the larger SE.

For the combined Dataset 1 comprising all animals, when the pedigree-free de-regressed EBVs were used for Population 2, the genomic heritability was *0.14 (SE = 0.05)*, and for the pedigree-derived de-regressed EBVs it was *0.11 (SE = 0.04)*. For the combined Dataset 2 after removing the lower reliability bulls, the genomic heritability was *0.19 (SE = 0.06)* (Table 2). These values are indicative of genetic variation but care should be taken in their interpretation due to the different trait definitions, hence expected values of the traits, in the different populations and datasets.

4.3.2 Regional heritability estimates

Windows contributing the maximum heritability and with the maximum LRT test value were identified for each chromosome, for the three window sizes. For the combined Dataset 1 and when the pedigree-free de-regressed EBVs were used for Population 2, RH_{max} ranged from *0.005* to *0.032* for the 20-SNP window, from *0.004* to *0.029* for the 30-SNP window, and from *0.006* to *0.028* for the 50-SNP window (Table 3). When pedigree-derived de-regressed EBVs were used for Population 2, RH_{max} ranged from *0.172* to *0.409*, from *0.009* to *0.359*, and from *0.007* to *0.353* for the *20*, *30* and *50*-SNP windows (Appendix 4.1). However, as demonstrated by Ekine et al. (2013), inclusion of pedigree information is likely to inflate the estimates.

The strongest evidence for association was on chromosome 6 (BTA6) for the 50-SNP window (Fig. 4a and b), significant at the suggestive significance threshold *(LRT_{max}= 9.19, suggestive threshold= 8.96)* when using the pedigree-free deregressed EBVs for Population 2, with the window (position: 45,216,251-48,752,176, Fig. 5) explaining *2.7%* of the phenotypic variance (Tables 3 and 4). This result was not observed by either Bermingham et al. (2014) or Finlay et al. (2012). With pedigree-derived de-regressed EBVs, BTA6 gave an LRT_{max}= *7.16* for the 50-SNP window and h^2_r =0.238, with this value again likely to be an overestimate (see Appendix 4.1). LRT results from the RH mapping analysis on Population 2 alone are shown in Appendix 4.2.

For completeness, RHM was repeated for Dataset 1, but with the difference that the same quality control criteria were applied to both populations. In this analysis the strongest evidence for association was on BTA6 for the 50-SNP window (Fig. 4c and d), significant at the suggestive significance threshold (LRT_{max}=*10.04*, suggestive threshold=*8.90*). The window (position: 44,698,534- 47,983,800), as found previously and shown in Fig. 5, explained *2.7%* of the phenotypic variance (Tables 5 and 6). Approximately the same region (position: 44,461,834 - 47,306,228) for the 30-SNP window was again significant at the suggestive level (LRT_{max}=*10.30*, suggestive threshold=*9.83*).

In Dataset 2, after removing the low reliability bulls, the strongest evidence for association was again on BTA6 for all the different window sizes tested, however, no region reached significance after the Bonferroni correction (see Appendix 4.3). RH_{max} ranged from *0.008* to *0.029* for the 20-SNP window, from *0.006* to *0.030* for the 30-SNP window, and from *0.006* to *0.039* for the 50-SNP

window (see Appendix 4.3).

When the two populations in Dataset 2 were re-analysed, but setting the covariance between the two populations in the **G** matrix to zero, BTA3 and BTA14 provided high LRT values while for BTA6 the LRT was reduced compared the previous analysis (see Appendix 4.4).

4.3.3 GWA analysis

GWA analysis on all the animals and after taking into account possible population sub-structure through fitting principal components, did not yield any significant associations at the chromosome or genome-wide levels ($-log_{10}(P\text{-value}) = 4.24$. The most significant SNP identified was compared to the genome-wide and suggestive significance thresholds of *5.85* and *4.55* respectively (see Appendix 4.5).

4.3.4 Chromosomal heritability estimates

Chromosomal heritability estimates for the different methods described in 4.2.2.3 are presented in Table 7 for Dataset 2 (for Dataset 1 results see Appendix 4.6). Fitting chromosomes individually following Method (a) in 4.2.2.3, resulted in a gross overestimation of the individual chromosome and total heritability, compared to fitting them simultaneously (Fig. 6 and Fig. 9). Method (b), where all chromosomes are fitted simultaneously, corrects for the correlation of genotypes across chromosomes. Method (c), fitting a single $\mathbf{G}_{-c}$ for all remaining chromosomes, also takes into account that the chromosome under study might be correlated with other chromosomes and provides a simpler model for analysis. This approach can be

useful particularly in situations where there is not enough information available (e.g. very small sample size) for fitting a large number of components simultaneously as in Method (b). However, constraining all chromosomes to have the same pattern of variance may not be appropriate as for example large QTLs under selection may be segregating in some chromosomes but not in other chromosomes, and the pattern of relationship on one chromosome may be different to the pattern of relationship on another chromosome. Therefore, Method (b) provides the most reliable estimates and is considered to be the optimal approach for obtaining chromosomal heritability estimates.

There was a tendency for chromosomes which contained the most significant regions identified in 4.3.2, to also have the highest chromosomal heritability. All the methods (with the exception of Method (a)) were in good qualitative agreement providing the largest chromosomal heritability estimate for BTA6 (Fig. 9), confirming findings from the RH mapping, a. The chromosomal heritability estimates for BTA6 were *0.059*, *0.051*, *0.048*, *0.053* and *0.053* from Methods (a), (b), (c), (d) and (e) respectively. These estimates were consistent with a regional heritability of *0.027* for BTA6. The sum of the individual chromosome heritabilities is in the same range as the trait heritability: *0.284* for fitting all *29* chrs in Method (b), and *0.289* for fitting *20* non-zero variance chromosomes in Method (e), see Table 9. The chromosomes accounting for most of the observed variation were BTA3, BTA6 and BTA14, providing LRT values significant at the suggestive level (both the REML programme and ASReml provided similar log-likelihood estimates, with genome-wide significance threshold: *9.82* and suggestive threshold: *4.47*) (Table 8).

When the *20* chromosomes with non-zero variance in Method (b) were fitted simultaneously with the genomic matrix on the remaining chromosomes, these were found to explain *29%* of the total phenotypic variance, suggesting that there are a few major chromosomes affecting resistance to bTB ($\sigma^2_{20} / \sigma^2_P = 0.29$, where σ^2_{20} is the sum of the variance explained by the *20* chromosomes fitted simultaneously and σ^2_P the total phenotypic variance). In this analysis, for three of those chromosomes (BTA11, BTA20, and BTA25) the chromosomal heritability became zero, and only *17* of the *20* chromosomes contributed to σ^2_{20} (Table 9).

The genetic architecture of the trait can be investigated through the slope of the regression of $h^2_{c(sep)}$ on chromosome length. For highly polygenic traits and in the absence of population structure the proportion of variance explained by each chromosome (i.e. the variance captured by the **G** matrix calculated on each chromosome) is expected to be proportional to its length (Yang et al. 2011; Daetwyler et al. 2012). However, in both the analyses of Dataset 1 and Dataset 2, chromosomal heritability was not found to be related to chromosome length. The slope was not significantly different from zero, with *P-value>0.1* for Dataset 1 and *P-value>0.1* for Dataset 2. This, along with the substantial number of chromosomes with zero heritabilities, suggest that bTB resistance is not strictly infinitesimal, and is controlled by certain chromosomes (Fig. 7, Fig. 10). When the heritability was regressed on the chromosomal length fitting the SE as weighting factors, the intercept or regression slope did not change considerably.

The regression of the difference of the heritability estimates when (a) the chromosomes are fitted individually ($h^2_{(sep)}$), and when (b) are fitted simultaneously (h^2_c), i.e. $h^2_{(sep)} - h^2_c$, on the chromosomal length, is indicative of population structure

(i.e. relatedness) (Yang et al. 2011), and represents the impact of markers accounting for variation in the phenotype even though they are on different chromosomes from where the QTL may be. For both the analyses of Datasets 1 and 2, the slope of this regression was not significantly different than zero (*P-value > 0.01* for Dataset 2) indicating that there was no evidence for population stratification due to systematic differences in allele frequencies between the two subpopulations (Fig. 8, Fig. 11). The intercept of this regression is due to cryptic relatedness (Yang et al. 2011), and was found to be significantly different than zero (*P-value < 0.01*).

From the ratio of the intercepts b_0 and $b_{0(sep)}$ (Daetwyler et al. 2012), the proportion of genetic variance due to population structure was estimated to be *0.80* (*0.85* for Dataset 1) suggesting that *~80%* of estimated genomic h^2 is due to relatedness. These results suggest that although the markers capture some effects of individual loci through linkage disequilibrium (LD), additive pedigree-correlated relationships are likely to play an important role in the total genetic variation captured by the markers.

4.3.5 Genomic prediction

Cross Validation (CV) methods were used to test the accuracy of the within-population and across-populations genomic predictions (Luan et al. 2009). Increasing the size of the training set is expected to be beneficial in a cross validation. Therefore, the two populations were combined in a 5-fold cross validation and predictions within and across populations were compared.

The average prediction accuracies obtained on the combined populations across *50* repeats were $r(g,\hat{g})=0.33$ *(s.d. 0.05)* for the pedigree-free EBVs, and

r(g,ĝ)=0.38 (s.d. 0.05) for the pedigree based EBVs (Table 10 and Fig. 12), compared to *0.33* when Population 1 alone was analysed (Tsairidou et al., 2014). Consistent with results mentioned above, including pedigree information in the EBVs resulted in greater estimates. Across populations prediction, even when using EBVs with pedigree information, resulted in reduced accuracy *(r(g,ĝ)=0.1*), and when EBVs with no familial information were used the correlations for across population predictions were close to zero (Table 10). These results suggest that genomic prediction is feasible but less accurate when applied across disparate populations. A possible explanation for that is that the accuracy depends on the genetic relationship between the validation set and the training set, and there might be systematic differences in linkage phases across the populations. Prediction accuracy was improved when individuals from the population to be predicted were included in the training set.

When analysing the combined Dataset 2 with the low reliability bulls removed the average correlation with phenotype, obtained across *50* randomisations, was *0.14* and the average accuracy was *r(g,ĝ)=0.34 (s.d. 0.04)* (Table 10 and Fig. 13). For the cross validation across populations the correlation and accuracy were *0.011* and *0.03* when predicting for Population 2. However, the reverse prediction did not provide a credible value (Table 10).

4.4 Discussion

In the present analysis, where two disparate datasets were combined, it was demonstrated that bTB resistance is a moderately polygenic trait. A large effect was identified on BTA6 with most methods used being in good agreement. These results

confirm the analyses presented in earlier chapters that genomic prediction for bTB resistance is feasible, however for the present data across population prediction was found to be of little value. Exploration of the properties of the chromosomal heritabilities suggests that a high proportion of the predictive ability of the SNPs is due to additive genetic relatedness rather than markers closely tagging causal mutations.

4.4.1 Genomic heritability and regional heritability mapping

For Population 1 the genomic heritability estimate was *0.23 (SE = ±0.06)*, while for Population 2, retaining only the *175* bulls with ≥*8* daughters resulted in the genomic heritability estimate of *0.60 (SE = ±0.22)*. These values are indicative of genetic variation but care should be taken in their interpretation due to the different trait definitions, hence expected values of the traits, in the different populations and datasets. Population 1 is a case/control study while the Population 2 is a random sample of bulls. Bermingham et al. (2014) reported an estimated heritability of bTB resistance in Population 1 of *21.0%* (*95%* CI: *8.6–33.4*) on the observed scale. For Population 2, de-regressed EBVs estimated with no pedigree information are daughter averages corrected for fixed effects, and the heritability is essentially an estimate of the population average reliability (i.e. $n/(n+(4-h^2)/h^2)$) where n is the harmonic mean number of daughters and h^2 is the true trait heritability for a single phenotype. For a harmonic mean of *17.24* daughters per sire in this subset, this heritability of *0.60* implies a heritability of single phenotype $E[h^2]=0.32$.

For the combined analyses the genomic heritability estimates were *0.14 (SE = 0.05)* for Dataset 1 and *0.19 (SE = 0.06)* for Dataset 2, lower than those obtained

from individual populations. A possible explanation could be that assuming a covariance between the two populations in the IBS matrix added noise rather than information. To address this hypothesis the heritability analysis was repeated setting the across-population covariance in the **G** matrix to zero. This analysis yielded a heritability of *0.23 (SE = 0.06)*, similar to the estimate obtained when analysing Population 1 alone, suggesting that there is enough information in the SNPs present in the low density SNP chip and that when assuming no covariance between the two populations, Population 2 is outweighed due to its unreliable heritability estimate. Population 2 was a rather heterogeneous population with big differences in the reliabilities as derived from progeny testing in cattle, and therefore appears not to add information for the estimation of the heritability, which had a large standard error due to the small sample size. Therefore, the estimate of $h^2=0.23$ as yielded from Population 1 is likely to be closer to the true heritability for bTB resistance. Assuming no covariance between the two populations may be more appropriate for the purposes of estimating genomic heritabilities and particularly when populations are distant and marker effects are likely to be different in the two populations.

When assuming no across-population covariance, the LRT for BTA6 was reduced compared to the analysis presented in 4.2.2.1, while BTA3 and BTA14 provided high LRT values (see Appendix 4.4). This result suggests that for BTA6 there may be ancestral haplotypes common to both populations, which in this case allowed for a gain in power when combining the populations.

When each of the populations were analysed independently, associations could be identified on BTA13 for Population 1 as reported by Bermingham et al. (2014), while for Population 2 the QTL observed by Finlay et al. (2012) could not be

replicated in this analysis. GWA analysis on the combined populations did not detect any significant associations, which is an example of increased power of RH mapping to identify genomic regions associated with the trait under study compared to single-SNP approaches (Nagamine et al. 2012; Uemoto et al. 2013). However, allele frequencies and SNP-mutation linkage phases may also differ between populations.

4.4.2 Genomic architecture of resistance

The chromosomal heritability estimation analysis revealed an intermediate situation for the genetic architecture of bTB resistance, where the trait was not found to be highly polygenic while there were not shown any major gene associations. Resistance to bTB was found to be affected by major effects on a few chromosomes while relatedness was found to be also playing an important role. The utility of taking into account the genetic architecture of the trait under study has been highlighted before in the literature, and prediction accuracy might benefit from incorporating this knowledge and the distribution of QTL effects in the analysis as prior information (Hayes et al. 2010; Moser et al. 2015).

Fitting chromosomes individually following method (a) in 4.2.2.3, resulted in a gross overestimation of the individual chromosome and total heritability, compared to fitting them simultaneously (Fig. 6 and Fig. 9). One possible explanation is that this is due to correlations of SNPs on different chromosomes as the result of population structure (Yang et al. 2011). From the chromosomal heritability analysis *17* chromosomes were found to explain *29%* of the total phenotypic variance. To test the hypothesis that a similar number of chromosomes would be found to contribute non-zero variance even if the QTLs were randomly located across the chromosomes,

a false map was provided in the analysis so that the SNPs that passed QC had their positions randomised over all *29* chromosomes (using the *"sample"* function in *R*). IBS matrices were calculated for every (false) chromosome and they were fitted simultaneously (as in Method (b)). A greater number of iterations was required to achieve convergence than when using the true map. In this analysis, all chromosomes were found to be contributing the same amount of variation, with all chromosomes detecting a baseline variance. These results suggest that the variance in the original analysis is genuinely associated with those SNPs on those particular *17* chromosomes, with some chromosomes contributing near zero variance and a few chromosomes contributing considerably more.

4.4.3 Genomic prediction

Genomic prediction accuracy is largely affected by the sample size, but the genetic distance between the validation and the training populations is also important as the more distant the populations are, the greater the systematic differences in allele frequencies and linkage phase across populations will be. Genomic prediction was found to be feasible within populations but across population prediction resulted in reduced accuracy and sometimes resulted in negative values. This finding was consistent with the genomic heritability estimate being higher when the between population IBS covariances were set to zero. Using Population 1 (*n=1,151*), to predict for the smaller validation set (Population 2) was feasible; however, the achieved accuracy was very low. When a very small sample was used as the training set (i.e. Population 2 after removing the bulls with less than *8* daughters), it did not provide reliable accuracy estimates. Further, this estimate is corrected by the heritability which given the small size of the dataset, it will not be a precise estimate.

Thus, inferences on across-population predictions should be drawn with caution.

When using the 5-fold cross validation in the combined populations, the prediction accuracy was not significantly improved by the increased sample size, compared to the estimates from a 5-fold cross validation within Population 1 alone. However, although the two populations were only distantly related and thus predictions were made across more diverse populations, combining the two populations was not found to have a negative impact on the prediction accuracy and the prediction accuracy was maintained. Furthermore, including in the training set individuals from the validation population, has been reported before to improve the accuracy of across populations predictions (Riggio et al. 2014b). In the present study, when using the cross validation in the combined populations, the training sets contain individuals from both populations, and this approach provided improved accuracy ($r(g,\hat{g})=0.34$ compared to 0.03 when predicting from Population 1 alone).

4.4.4 Genomic region and candidate genes associated with bTB resistance

BTB is a chronic inflammatory disease of the lungs and the respiratory system with the macrophages being the primary target cells of the *M. bovis* bacteria and the innate immune response playing an important role in the outcome of infection. The region identified (position: 45,216,251-48,752,176 bp) through the RH mapping approach, contains a number of annotated genes. Based on their relevance regarding bTB, three putative candidate genes were identified: DHX15, SLC34A2, and ECSOD. These genes are involved in immune response and have been previously associated with disorders of the respiratory system and the lungs (Fig. 5).

Specifically, they have been previously linked to chronic respiratory disease and conditions of chronic inflammation of the airways, with mutations in these genes affecting normal lung function and response to infection. In the following paragraphs the specific candidate genes are presented in more detail.

The most promising candidate is the DHX15 gene which affects IFN production and innate immune response (Table 11). It is a member of the helicase family and codes for the pre-mRNA-splicing factor ATP-dependent RNA helicase DHX15 enzyme. It functions as a Pattern Recognition Receptor (PRRs), recognising and binding viral RNA (dsRNA), activating the immune response (myeloid dendritic cells) and inducing interferon production (Type I IFN), playing an important role in the innate immune system response (Lu et al. 2014). PRRs and their coding genes have been suggested in the literature as potential candidates for selection for disease resistance (Kaiser 2010, p. 16). Moreover, preliminary results from qPCR analysis studying the expression levels of this gene in BCG-activated macrophages, have provided indication that DHX15 is up-regulated (*2-3* fold up-regulation) throughout infection (Mühlbauer L., personal communication, July 8, 2015). A previous gene expression study showed that DHX15 was expressed *5.6* times more in *M. bovis* infected alveolar macrophages compared to *M. tuberculosis* infected bovine macrophages (Widdison et al. 2008).

Secondly, the Solute Carrier family 34 member 2 (SLC34A2) encodes the sodium-dependent phosphate transport protein 2B and is mainly expressed in the lungs (see Appendix 4.8), and specifically in the type II alveolar cells. Mutations in this gene that disrupt phosphate metabolism in the phospholipid degradation process by the type II cells and the alveolar macrophages, have been associated in humans

with excessive calcification in the alveolar space and an autosomal recessive disorder called Pulmonary Alveolar Microlithiasis (PAM) which leads to the development of chronic inflammation and chronic respiratory failure (Huqun et al. 2006). Moreover, another member of the same family, Solute Carrier family 6, member 6 (SLC6A6), has been previously associated with susceptibility to bTB in a previous study on data from Population 2 (Finlay et al. 2012). SLC6A6 codes for the sodium- and chloride-dependent Taurine Transporter (TauT) which is a carrier protein for the taurine amino acid across lipid layers (such as membranes). TauT deficiency has been linked to differences in immunological inflammatory response and response to bacterial infection. However, SLC6A6 is on a different chromosome (BTA22) and this result was not replicated in the present meta-analysis. Further, a third member of the solute carrier family is SLC11A1 gene on BTA2, encoding the Natural Resistance Associated Macrophage Protein 1 (NRAMP1), is expressed in lysosomes of monocytes and macrophages and is involved in the activation of macrophages and the innate immune response. Although in the present study it was not found to be significant, in previous studies polymorphisms at four loci in SLC11A1 have been associated with susceptibility to the *M. tuberculosis* (Li et al. 2011). Further NRAMP has been associated with susceptibility to infection with *M. bovis* BCG in mice (Vidal et al. 1993), and with the within-macrophages survival of *M. bovis* BCG, *Brucella abortus* and *Salmonella dublin* in cattle (Qureshi et al. 1996; Allen et al. 2010).

The third candidate, the ECSOD or SOD3 gene, is a member of the Superoxide Dismutase multigene family and codes for the Extracellular Superoxide Dismutase enzyme. SOD3 is expressed extracellularly in lung tissue in high

concentrations, playing an important role in antioxidant defence and functioning as an anti-inflammatory (anti-oxidant) protein protecting the lung from oxidative stress. Polymorphisms in this gene have been shown to be important in lung function and were previously associated with lung diseases such as the Chronic Obstructive Pulmonary disease (COPD) where ECSOD polymorphisms have been identified as COPD genetic risk factors (Wilk et al. 2007; Oberlay-Deegan et al. 2009). COPD is a persistent chronic inflammatory disease of the airways and its pathology is linked to the production of Reactive Oxygen Species (ROS) by the inflammatory cells e.g. macrophages (free radicals production) (Oberlay-Deegan et al. 2009). Furthermore, SOD3 is a copper-enzyme. Cu deficiency and oxidative stress on the erythrocytes have been linked to increased Heinz body haemolytic anaemia (Suttle et al. 1987), while genetic differences in copper metabolism have been previously associated with growth retardation and increased susceptibility to infection (Woolliams et al. 1986a and b).

Lastly, this region identified through the regional heritability mapping approach is ~*35* Mb away from the SNP identified on BTA6 from the heterozygote disadvantage analysis for bTB resistance presented in Chapter 3. However, that was an analysis capturing non-additive genetic variation and therefore, it was not surprising that that SNP was not detected in the present analysis.

4.4.5 Phenotypes and breed definitions

The animals comprising the populations under study were designated as Holsteins or Friesians (for Population 1 see Chapter 1 section 2.2.1). The presence of both Holsteins and Friesians in the datasets analysed was not considered to affect the

results since when Population 1 was analysed separately with the Friesian cows removed only minor effects could be observed on the heritability and accuracy estimates (h^2*=0.21 (SE = 0.07)* and *r=0.36 (95% C.I.: 0.33, 0.38*), Tsairidou et al. 2014). Moreover, the principal component analysis on the combined data did not show any substructure attributable to Holstein-Friesian breed differences, while the presence of beef cattle crossbreds was presented in the PCA and addressed by correcting through the **G** matrix and by fitting principal components as fixed effects in the models used.

In the present study we have combined individual phenotypic performance records (i.e. case / control data) and EBVs calculated from half-sib offspring averages (i.e. mean of individual measurements on several daughters). EBVs may be expected to improve the prediction accuracy as a "more accurate phenotype" as they include information from multiple progeny from each sire, hence using EBVs corresponds to the use of a trait with higher heritability (given that there are enough daughters, reliability will be such that $r^2 > h^2$). However, one of the direct consequences of using BLUP (Best Linear Unbiased Prediction) in the calculation of the EBVs is that the EBVs are shrunk towards the mean depending on the amount of information available (i.e. their reliability). The reliability is an estimate of the squared correlation between the EBV ($\hat{g}$) and the unknown True Breeding Value (TBVs, g), and it inversely reflects the prediction error variance associated with the estimation of the EBVs (so that $\hat{g} = g + (\hat{g} - g)$, where $(\hat{g} - g)$ is the prediction error). The EBVs estimated through BLUP have smaller variance compared to the TBVs so that $var(\hat{g}_i) < var(g_i)$, and that is due to the shrinkage towards the mean which depends on the amount of information available, i.e. the number of progeny. For

truly superior animals, EBVs from BLUP are lower than their TBVs, and vice versa for truly inferior animals. Therefore, the use of deregressed EBVs (i.e. EBVs divided by their corresponding reliability $\hat{g} / r^2_i$) has been suggested in the literature as showing advantages over the use of EBVs specifically when the datasets comprise EBVs for individuals with varying r^2_i (Garrick et al. 2009; Ostersen et al. 2011). Secondly, Garrick et al. (2009) point to the problem of deregressed EBVs having an excess weighing towards the parent average. In the present study, the influences of parent average effects were accounted for through the use of pedigree-free deregressed EBVs, which were calculated in the absence of ancestral information and thus, they did not have excess weighting towards any parental information.

Garrick et al. (2009) proposed weighting the de-regressed EBVs in order to correct for their heterogeneous variances. Weights can be calculated as a function of individual animal reliability (r_i^2) and the proportion of genetic variance that is not accounted for by the markers (c), (Equation [10] in Garrick et al. 2009). A value suggested for c in the literature is that of *0.1* from a study on a pure-bred pig population using the Illumina PorcineSNP60 BeadChip (Ostersen et al. 2011). Assuming a $c=0.2$ corresponding to the Bovine50 SNP chip (Daetwyler 2009, Thesis Chapter 7, p. 152) and using the individual bulls' reliabilities and $h^2=0.23$, the weights calculated for the bulls in Population 2 after removing those with <8 daughters, were found to range from *0.29* to *10.79* (with a mean of *1.34*). For the cows in Population 1, with $c=0.2$ and $r^2=h^2=0.23$, following the same formula, the weight can be calculated as *0.94*. Therefore, in a weighted regression analysis using these weights, the cows in Population 1 would get a considerably lower weight compared to the bulls in Population 2. Information on c was not available for this

dataset and given that two different SNP chips were used for the two populations, only assumptions could be made for the value of *c* in the combined populations. The approach used in the present study, of scaling the phenotypes to have equivalent standard deviations, may have resulted in over-weighting the de-regressed EBVs from Population 2, but that should not prevent identification of genuine associations.

4.4.6 Conclusion

Combining two independent datasets has provided insights into the inheritance of bTB resistance, from the level of the genome down to potential candidate genes. Interrogation of chromosomal heritabilities suggests that a high proportion of the prediction accuracy is due to relatedness between animals. Resistance to bTB was found to be a moderately polygenic trait. Loci which affect resistance are spread across a number of chromosomes but, critically, not all chromosomes, suggesting that there are a few major chromosomes affecting the trait. The most significant individual region identified is on BTA6, a region containing several plausible candidate genes that are involved in immune response and affect the function of the respiratory system, potentially affecting resistance to infection. The approaches used in the present study suggest that this chromosome has an important contribution to the observed variation in this data. This result was obtained through the meta-analysis of the two populations and it was not evident from the individual studies. Genomic prediction of bTB resistance in cattle, and in the absence of animal phenotypes or pedigree information, does appear to be feasible, even when the genotypes available have been obtained using lower density SNP platforms. However, across population prediction was not successful due to the small size of the training sets, and also most likely due to the genetic distance between the two

populations. Further, despite the common purpose of the phenotypes used, they were not defined precisely in the same way. Before implementation, the prediction accuracy needs to be improved, and particularly in situations where the populations may only be distantly related and where we wish to draw inferences on across-populations predictions.

	n animals	n SNPs	n SNPs after QC
Population 1	1151	777962	588332
Population 2	287	54001	41418
Population 2 (≥8 d)	175	54001	41428
Combine dataset (all animals) QC1	1438	-	37398 (36690 autosomal)
Combine dataset (all animals) QC2	1438	-	35459 (35021 autosomal)
Combined dataset (≥8 d Pop. 2)	1326	-	35427 (34987 autosomal)

Table 1. Number of animals and number of SNPs before and after Quality Control (QC), for each population (upper part of the table). Number of animals and number of SNPs for the combined datasets (lower part of the table).

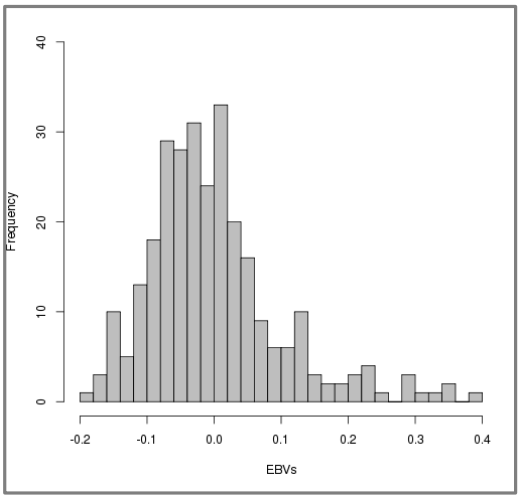


Figure 1. Histogram of pedigree-free de-regressed EBVs. Higher EBVs correspond to higher mean susceptibility, i.e. bulls with daughters of higher incidence rate.

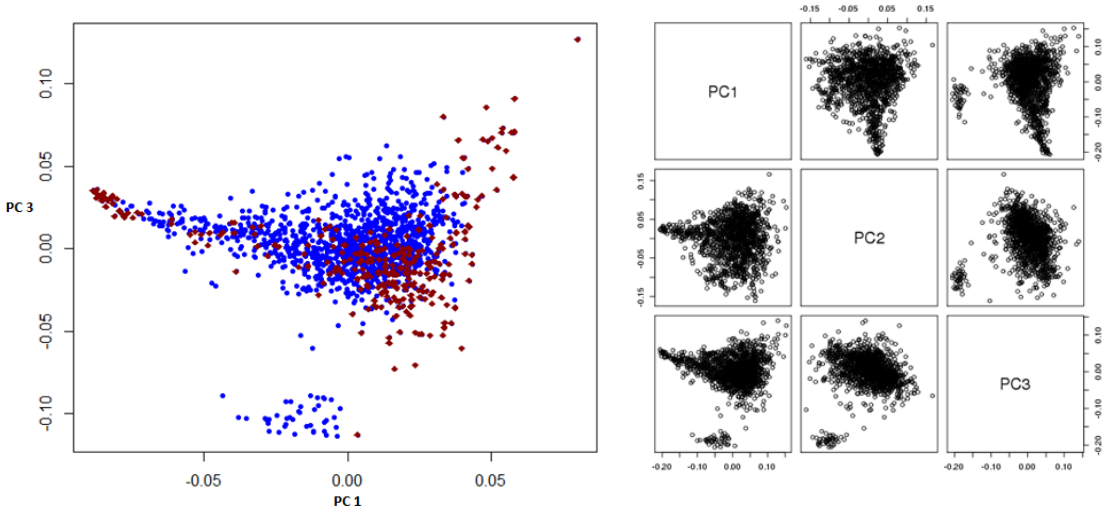


Figure 2. Principal component analysis for the combined populations *(n=1,438)*. In the graph are plotted the first vs. the third Principal Components (PC). The blue dots represent the Holstein-Friesian cows of Population 1, as presented in the structure graph by Bermingham et al. (2014). The red dots represent the Holstein-Friesian bulls of Population 2 as presented in the structure graph by Finlay et al. (2012). The graph on the left side shows the structure observed when plotting the combinations for the three larger principal components.

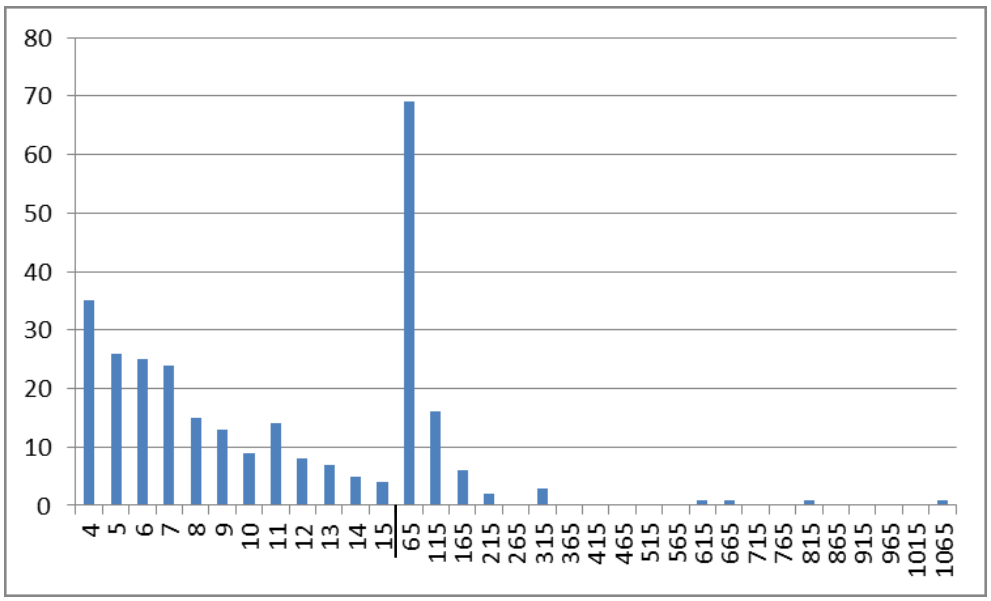


Figure 3. Distribution of number of daughters per bull in Population 2.

Dataset	Analysis	Pedigree for Population 2	n	Genomic h^2 (SE)
Population 1		-	1151	0.23 (0.06)
Population 2	EBV/ r^2_i	No	287	0.00 (-)
	EBV/ r^2_i	Yes	287	0.11 (0.10)
Population 2 (≥8 d)	EBV/ r^2_i	No	175	0.60 (0.22)
Combine dataset (all animals)	EBV/ r^2_i	No	1438	0.14 (0.05)
	EBV/ r^2_i	Yes	1438	0.11 (0.04)
Combined dataset (≥8 d Pop. 2)	EBV/ r^2_i	No	1326	0.19 (0.06)

Table 2. Genomic heritability estimates for the two populations and for each of the combined datasets and for the different sets of phenotypes analysed for Population 2 (where *n* is the number of animals in the analysis and r^2_i is the reliability of the i^{th} bull).

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	0.013	0.013	0.012
2	0.010	0.017	0.010
3	0.023	0.019	0.027
4	0.014	0.013	0.017
5	0.017	**0.029**	0.011
6	0.026	0.022	0.027
7	0.008	0.008	0.009
8	0.016	0.016	0.014
9	0.005	0.004	0.006
10	0.012	0.009	0.006
11	0.011	0.013	0.008
12	0.023	0.018	0.022
13	0.015	0.021	0.011
14	0.021	0.025	**0.028**
15	0.018	0.017	0.012
16	0.015	0.015	0.015
17	0.016	0.016	0.026
18	0.015	0.013	0.015
19	0.013	0.012	0.014
20	0.011	0.011	0.010
21	0.013	0.008	0.008
22	0.017	0.020	0.023
23	0.011	0.011	0.014
24	0.012	0.021	0.016
25	0.011	0.007	0.008
26	0.013	0.011	0.008
27	0.014	0.020	0.015
28	0.011	0.011	0.010
29	**0.032**	0.017	0.015

Table 3. Regional maximum fraction of phenotypic variance explained within each chromosome (RH_{max}) for Dataset 1 (*n=1438*), for the three window sizes tested. For each window size the maximum value is in bold. As phenotypes were used the pre-corrected residuals for Population 1 and the pedigree-free de-regressed EBVs for Population 2.

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	5.64	5.64	5.64
2	2.89	3.20	1.37
3	6.95	8.01	8.41
4	2.76	2.42	3.37
5	6.65	3.46	2.16
6	9.08	7.51	**9.19**
7	3.39	2.52	2.42
8	6.26	5.80	4.32
9	1.27	1.14	1.30
10	4.12	1.63	1.25
11	2.65	2.08	1.60
12	7.59	5.92	5.27
13	3.61	2.88	2.18
14	5.99	5.84	5.26
15	6.78	5.57	3.68
16	3.51	2.25	3.26
17	5.22	4.85	4.51
18	6.38	6.20	6.55
19	4.28	3.45	3.07
20	4.21	4.06	1.93
21	1.92	1.07	0.89
22	9.97	7.07	6.39
23	2.99	3.31	2.03
24	1.82	1.94	0.81
25	2.81	1.80	1.71
26	1.78	1.48	1.30
27	5.33	6.27	4.86
28	5.33	5.18	3.84
29	4.77	4.19	2.89
Genome-wide threshold	16.26	15.50	14.53
Suggestive threshold	10.65	9.90	8.96

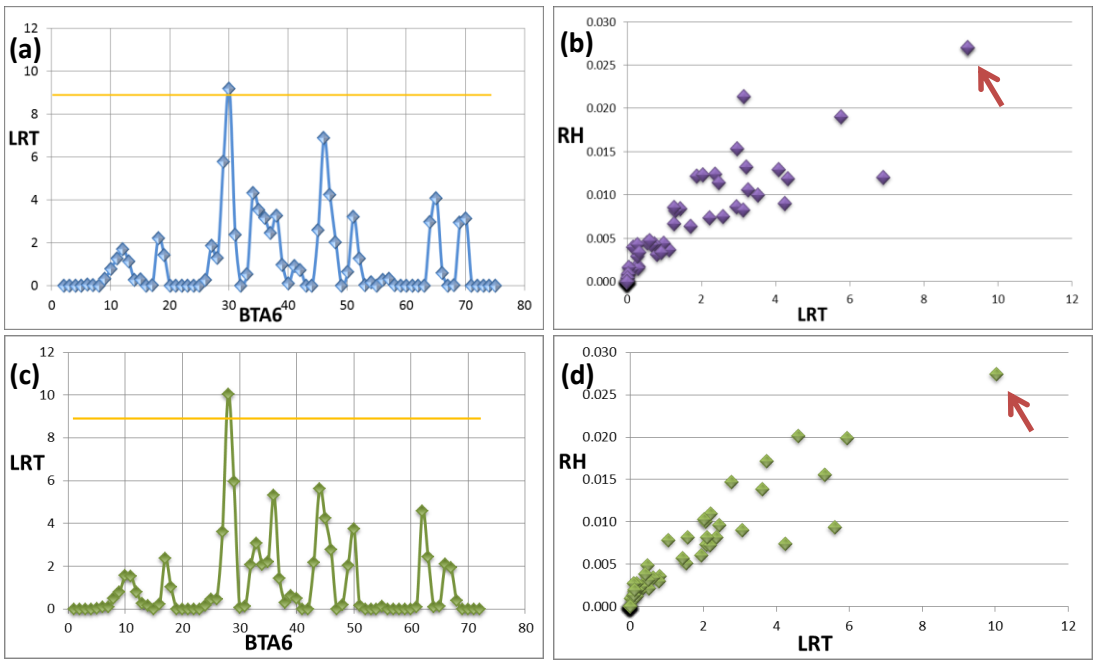
Table 4. LRT_{MAX} results within each chromosome for Dataset 1, and LRT significance thresholds after the Bonferroni correction, for the three window sizes tested. Significant value at the suggestive level is in bold. As phenotypes were used the pre-corrected residuals for Population 1 and the pedigree-free de-regressed EBVs for Population 2.

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	0.014	0.011	0.011
2	0.008	0.012	0.006
3	0.024	0.019	0.019
4	0.017	0.019	0.017
5	0.012	0.008	0.008
6	0.026	**0.028**	**0.027**
7	0.009	0.006	0.006
8	0.012	0.015	0.015
9	0.004	0.007	0.005
10	0.012	0.009	0.007
11	0.016	0.013	0.011
12	0.020	0.022	0.024
13	0.011	0.014	0.019
14	0.023	0.024	0.021
15	0.015	0.018	0.013
16	0.015	0.014	0.016
17	0.014	0.019	0.023
18	0.013	0.014	0.015
19	0.011	0.010	0.012
20	0.013	0.013	0.012
21	0.007	0.005	0.004
22	0.022	0.022	0.026
23	0.012	0.008	0.023
24	0.010	0.007	0.011
25	0.008	0.006	0.006
26	0.009	0.012	0.006
27	0.014	0.013	0.016
28	0.010	0.011	0.010
29	**0.031**	0.021	0.013

Table 5. RH_{max} estimates within each chromosome for Dataset 1 after applying the same QC criteria to both samples, for the three window sizes tested. For each window size the maximum value is in bold.

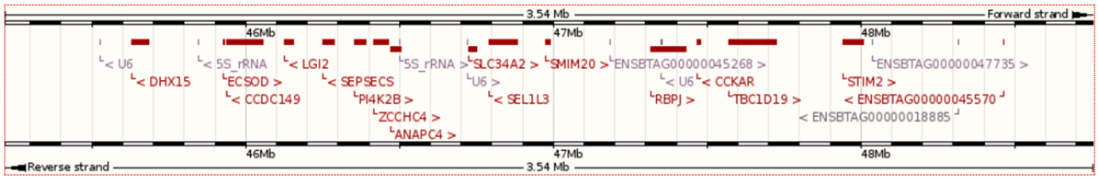
Window size and step size (number of SNPs)			
Chr	**20, 10**	**30, 15**	**50, 25**
1	6.11	3.56	2.17
2	3.22	2.29	1.35
3	8.65	8.86	8.31
4	3.72	2.79	2.79
5	3.76	1.83	2.16
6	9.14	**10.30**	**10.04**
7	2.99	2.27	1.49
8	5.15	5.89	4.08
9	1.06	1.82	1.03
10	2.83	1.75	1.31
11	2.29	2.05	1.94
12	8.37	7.56	6.69
13	2.44	3.02	2.79
14	6.20	6.03	5.18
15	4.49	5.81	3.99
16	2.96	4.61	3.94
17	5.49	4.69	3.62
18	5.88	5.39	5.49
19	3.87	2.34	2.20
20	5.14	4.25	2.96
21	1.07	0.93	0.44
22	8.64	6.64	5.35
23	3.15	1.66	4.33
24	1.68	0.89	0.96
25	1.83	1.89	1.51
26	2.25	1.71	1.33
27	5.51	7.12	4.08
28	4.40	4.36	3.97
29	4.02	4.47	2.48
Genome-wide threshold	16.19	15.43	14.46
Suggestive threshold	10.58	9.83	8.90

Table 6. LRT_{MAX} results within each chromosome for Dataset 1 after applying the same QC criteria to both samples and LRT significance thresholds after the Bonferroni correction, for the three window sizes tested. Significant values at the suggestive level are in bold.



*The yellow line represents the suggestive significance threshold.

Figure 4. Detailed analysis for Regional Heritability (RH) mapping for chromosome 6 (BTA6): (a) Log-likelihood Ratio Test (LRT) results for BTA6 from the 50-SNP window analysis, on Dataset 1 after QC1, with pedigree-free de-regressed EBVs for Population 2; (b) Quality control graph of the RH vs. the LRT values for all the regions on BTA6, where the arrow denotes the significant region. LRT and RH increase together as expected. (c) LRT results for BTA6 from the 50-SNP window analysis, on Dataset 1 after QC2; (d) Quality control graph of the RH vs. the LRT values for all the regions on BTA6, where the arrow denotes the significant region.



Analysis	Chr	Start SNP position (bp)	End SNP position (bp)
50-SNP window size (25-SNP step)	BTA6	Hapmap32456-BTC-038385 (45,216,251)	rs43459400 (48,752,176)

Figure 5. The region identified from the RH mapping analysis on Dataset 1, for the 50-SNPs window size, and the start SNPs and end SNPs of the region (http://www.ensembl.org/index.html).

Chr	Length(Mbp)	(a) $h^2_{c(sep)}$	(b) h^2_c	(c) $h^2_{(ci+G-ci)}$	$h^2_{c(sep)} - h^2_c$
1	161	0.000	0.003	0.000	-0.003
2	141	0.009	0.000	0.000	0.009
3	128	0.044	0.035	0.036	0.009
4	124	0.028	0.008	0.010	0.020
5	126	0.032	0.003	0.007	0.029
6	123	0.059	**0.051**	**0.048**	**0.008**
7	112	0.036	0.011	0.015	0.025
8	117	0.032	0.023	0.020	0.009
9	108	0.011	0.000	0.000	0.011
10	106	0.000	0.000	0.000	0.000
11	110	0.010	0.001	0.000	0.009
12	85	0.012	0.000	0.000	0.012
13	84	0.022	0.000	0.000	0.022
14	81	**0.061**	0.037	0.047	0.024
15	85	0.033	0.013	0.021	0.019
16	78	0.026	0.004	0.003	0.021
17	77	0.055	0.030	0.025	0.025
18	66	0.025	0.003	0.000	0.022
19	65	0.008	0.000	0.000	0.008
20	76	0.011	0.001	0.000	0.010
21	69	0.002	0.000	0.000	0.002
22	62	0.031	0.021	0.023	0.010
23	53	0.035	0.018	0.020	0.017
24	65	0.007	0.000	0.000	0.007
25	44	0.014	0.001	0.003	0.012
26	52	0.013	0.003	0.002	0.010
27	49	0.000	0.001	0.000	-0.001
28	46	0.022	0.016	0.015	0.006
29	52	0.023	0.000	0.004	0.023

Table 7. Chromosomal heritability estimation on Dataset 2 from Method (a) where every chromosome was fitted one by one, Method (b) where all the chromosomes were fitted simultaneously, and Method (c) where every chromosome was fitted one by one with the $\mathbf{G}_{-c}$ matrix calculated on the remaining chromosomes. In the last column are shown the values of the difference of the heritability estimates when each chromosome was fitted one by one and when all chromosomes were fitted simultaneously.

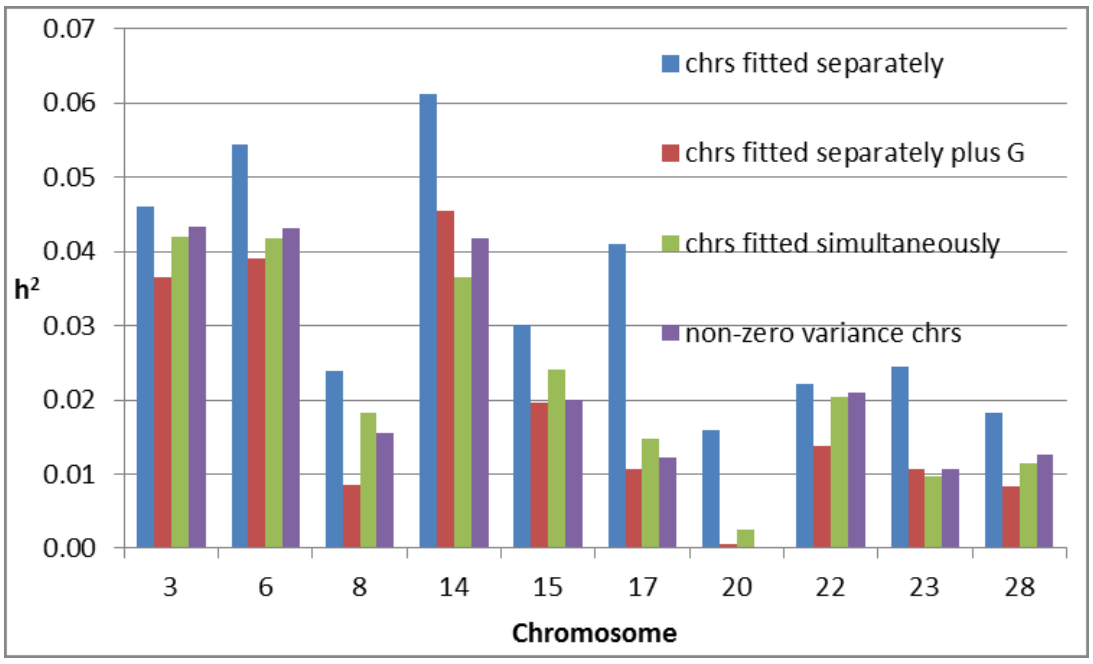


Figure 6. Chromosomal heritability estimates for Dataset 1, for the *10* chromosomes which had a non-zero variance in Methods (a), (b) and (c). The estimates shown are from the following analyses: Method (a) where heritabilities were calculated separately for each chromosome; Method (c) where each chromosome was fitted separately plus the whole **G** matrix; Method (b) where all *29* chromosomes were fitted simultaneously, and Method (f) where the chromosomes with non-zero variance in Methods (a), (b) and (c) were fitted simultaneously.

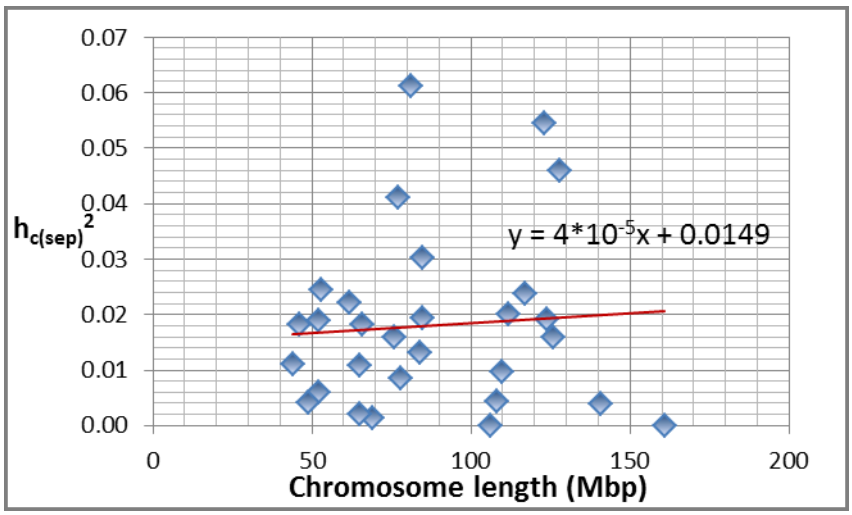


Figure 7. Regression of the heritability estimates calculated separately for each chromosome on the chromosome length, for the combined analysis of Dataset 1 ($b_{0(sep)}$= *0.0149,* $b_{1(sep)}$= *0.00004,* R^2 *=0.0053, P-value: >0.1*).

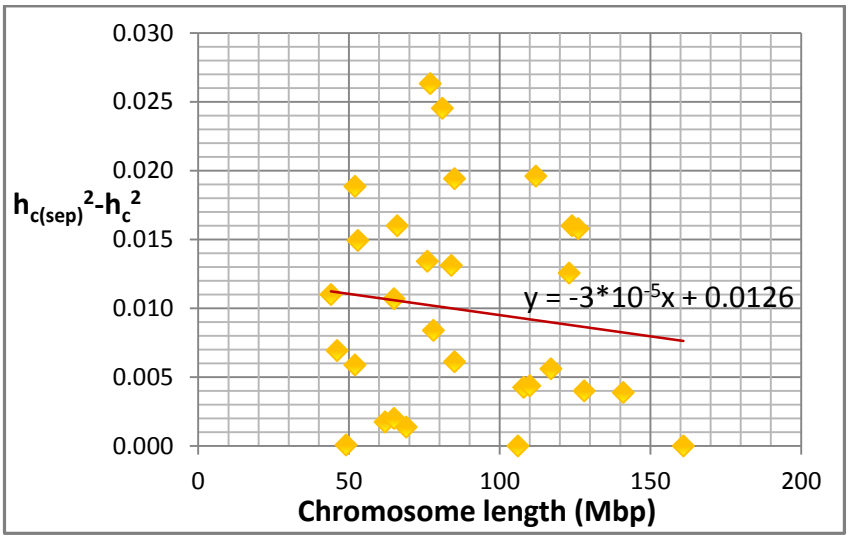


Figure 8. Regression of the difference between the heritability when the chromosomes were fitted individually in Method (a) and when were fitted simultaneously in Method (b) on the chromosomal length, for the combined analysis of Dataset 1 (*b_0 = 0.0126, b_1 = -0.00003, R^2 = 0.0169, P-value: >0.1*).

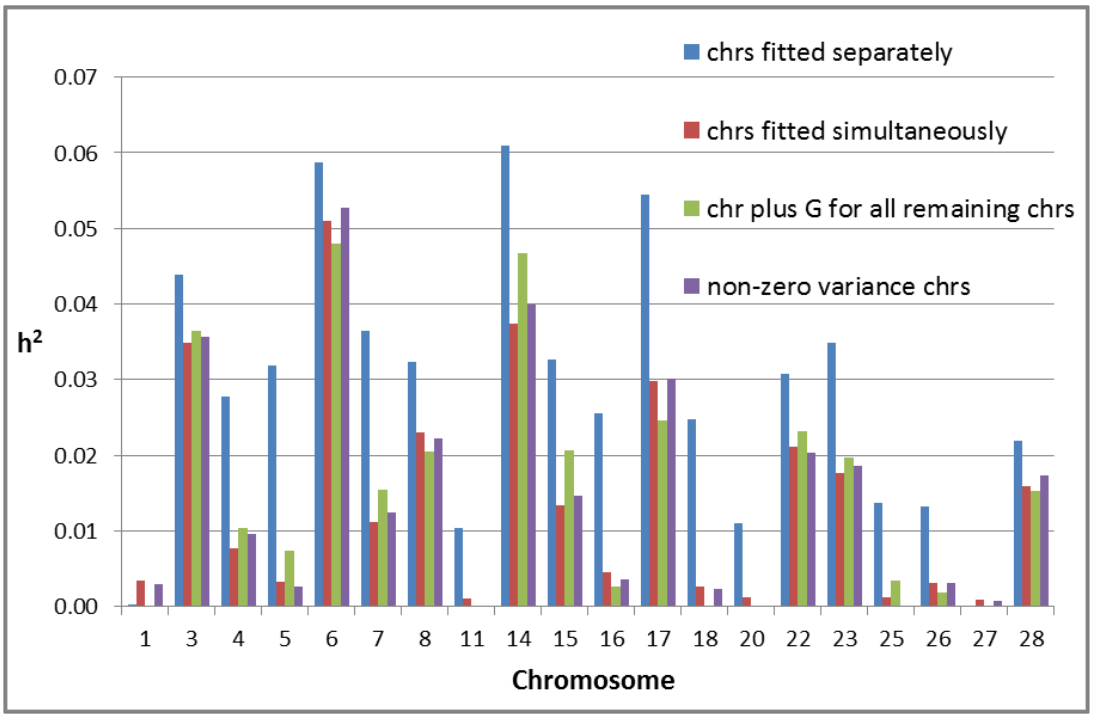


Figure 9. Chromosomal heritability estimates for Dataset 2, for the *20* chromosomes which had a non-zero variance in Method (b) when analysed using the model: $y = \beta + \sum g_c + e$. The estimates shown are from the following analyses: Method (a) when every chromosome was fitted one by one; Method (b) where all the *29* chromosomes were fitted simultaneously; Method (c) where every chromosome was fitted one by one with the $\mathbf{G}_{-c}$ on the remaining chromosomes; Method (e) where only the non-zero variance chromosomes in Method (b) were fitted with the $\mathbf{G}_{-c}$ on the remaining chromosomes.

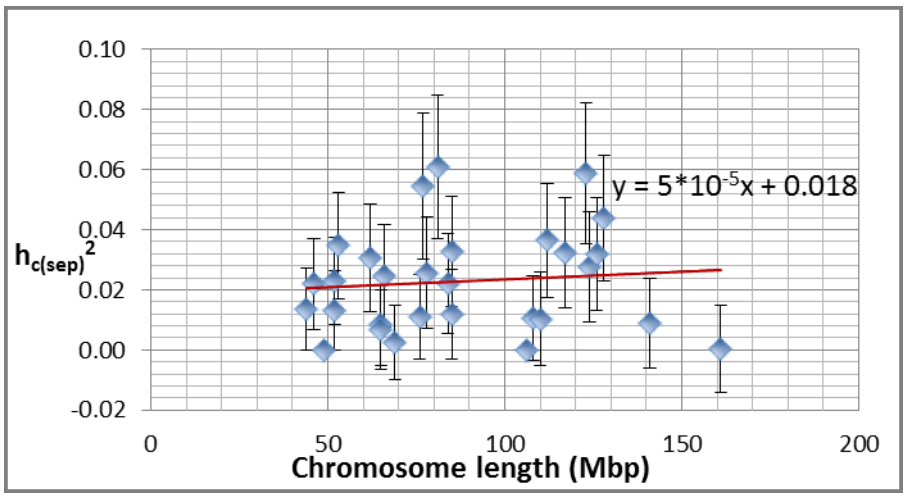


Figure 10. Regression of the heritability estimates calculated separately for each chromosome on the chromosome length with corresponding SE, for the combined analysis of Dataset 2 ($b_{0(sep)}$= *0.018,* $b_{1(sep)}$= *0.00005,* R^2 = *0.009, P-value: >0.1*).

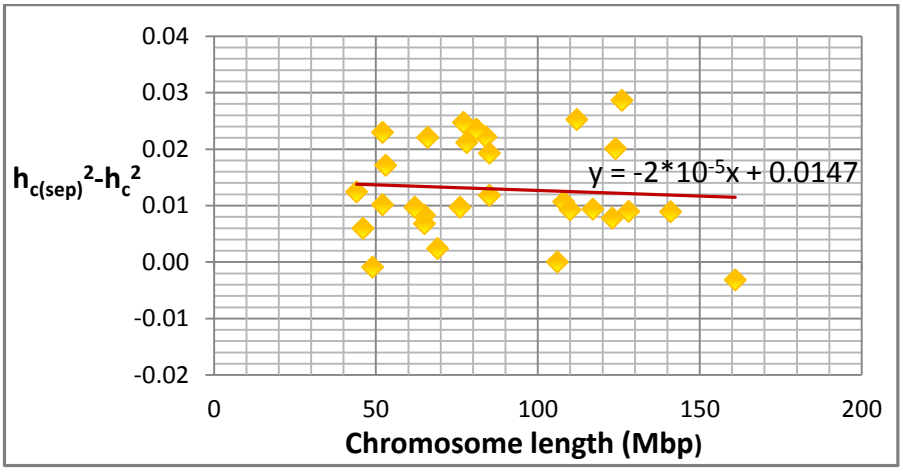


Figure 11. Regression of the difference between the heritability when the chromosomes were fitted individually in Method (a) and when were fitted simultaneously in Method (b) on the chromosomal length, for the combined analysis of Dataset 2 *(b_0 = 0.0145, b_1 = -0.00002, R^2 = 0.005, P-value: >0.1)*.

Chr	Chr_i component	Genomic-chr_i component	LRT AIREML
1	0.000	0.195	-0.049
2	0.000	0.212	-0.220
3	0.037	0.159	**5.795**
4	0.010	0.178	0.536
5	0.007	0.182	0.237
6	0.048	0.145	**7.829**
7	0.015	0.172	0.980
8	0.021	0.169	2.066
9	0.000	0.213	-0.237
10	0.000	0.222	-1.626
11	0.000	0.200	-0.045
12	0.000	0.206	-0.228
13	0.000	0.191	0.000
14	0.047	0.145	**8.681**
15	0.021	0.167	2.560
16	0.003	0.186	0.029
17	0.025	0.167	1.720
18	0.000	0.190	0.000
19	0.000	0.195	-0.038
20	0.000	0.188	-0.001
21	0.000	0.208	-0.293
22	0.023	0.175	2.895
23	0.020	0.163	2.132
24	0.000	0.204	-0.131
25	0.003	0.185	0.093
26	0.002	0.186	0.031
27	0.000	0.186	0.000
28	0.015	0.179	1.837
29	0.004	0.184	0.173
Genome-wide threshold			9.82
Suggestive threshold			4.47

Table 8. LRT_{MAX} results for Dataset 2 when fitting each chromosome separately and the $\mathbf{G}_{-c}$ on the remaining chromosomes with AIREML. The hypothesis that the chromosome is contributing variance to the trait is tested versus the H_0 model where only the genomic matrix excluding the chromosome under study is used.

	Method (b)		Method (e)	
Chr	**V_i**	**h^2**	**V_i**	**h^2**
chr 1	0.004	0.003	0.003	0.003
chr 3	0.040	0.035	0.041	0.036
chr 4	0.009	0.008	0.011	0.010
chr 5	0.004	0.003	0.003	0.003
chr 6	**0.059**	**0.051**	**0.060**	**0.053**
chr 7	0.013	0.011	0.014	0.012
chr 8	0.027	0.023	0.025	0.022
chr 11	0.001	0.001	0.000	0.000
chr 14	0.043	0.037	0.046	0.040
chr 15	0.016	0.013	0.017	0.015
chr 16	0.005	0.004	0.004	0.004
chr 17	0.034	0.030	0.034	0.030
chr 18	0.003	0.003	0.003	0.002
chr 20	0.001	0.001	0.000	0.000
chr 22	0.024	0.021	0.023	0.020
chr 23	0.020	0.018	0.021	0.019
chr 25	0.001	0.001	0.000	0.000
chr 26	0.004	0.003	0.003	0.003
chr 27	0.001	0.001	0.001	0.001
chr 28	0.018	0.016	0.020	0.017
V_{env}	0.828		0.812	
V_P	1.157		**1.142**	
V_{20}	0.329		**0.330**	

Table 9. Variance components (V_i) and corresponding heritability estimates when all chromosomes were fitted simultaneously in Method (b) and when only the chromosomes with non-zero variance in (b) were fitted together with the $\mathbf{G}_{-c}$ matrix on the remaining chromosomes (Method e). *17* chromosomes are found to explain *29%* of the total phenotypic variance, where V_{20} is the sum of the variance explained by those chromosomes. Three chromosomes, although they had a non-zero variance in the *29* chromosomes simultaneous analysis, their variance became zero in the *20* chromosomes analysis.

Dataset 1 CV	Training Population	Validation Population	h^2	SE	Pedigree for Population 2	Cor	Accur	SD
Combined analysis	1 + 2	1 + 2	ǂ		No	0.12	0.33	0.05
	1 + 2	1 + 2	ǂ		Yes	0.12	0.38	0.05
Across-population analysis	1	2	0.20	0.06	No	-0.01		
	1	2	0.20	0.06	Yes	0.04	0.10	
	2	1	0.00		No	-0.03		
	2	1	0.00		Yes	0.02		

Dataset 2 CV	Training Population	Validation Population	h^2	SE	Pedigree for Population 2	Cor	Accur	SD
Combined analysis	1 + 2	1 + 2	ǂ		No	0.14	0.34	0.04
Across-Population analysis	1	2	0.20	0.06	No	0.01	0.03	
	2	1	0.57	0.22	No	0.84	1.11	

ǂ The within-fold heritability for each Cross Validation fold was used

Table 10. Average correlation and expected accuracy across *50* randomisations from analysis of Dataset 1 and Dataset 2, and from the across population prediction. The values in red are non-informative values.

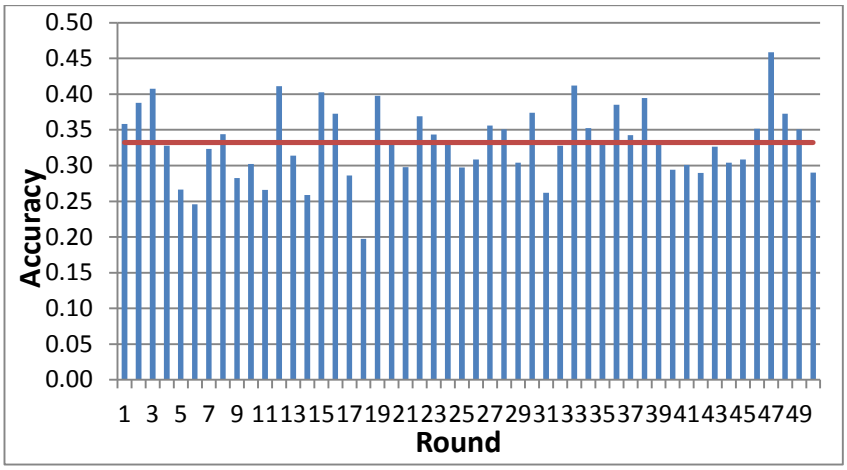


Figure 12. Accuracy values calculated as the mean of the *5* Cross Validation sets, across *50* randomisations, on Dataset 1. The red line denotes the mean value $r(g,\hat{g})=0.33$ *(s.d. 0.05)* as given in Table 10.

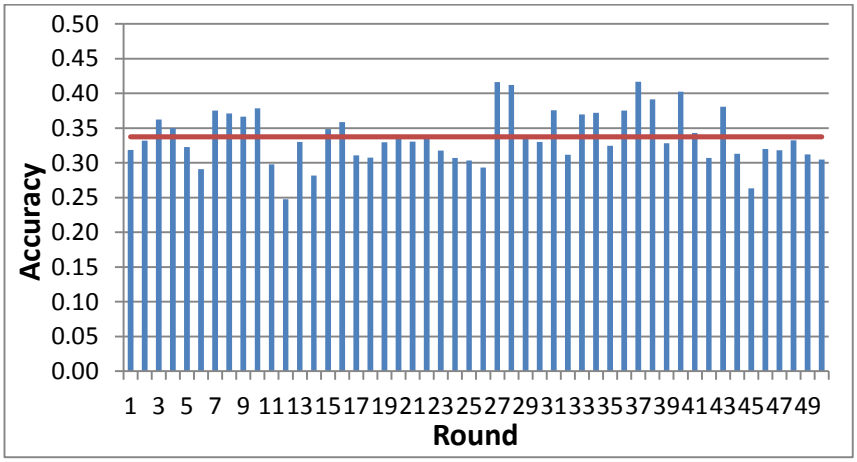


Figure 13. Accuracy values calculated as the mean of the *5* Cross Validation sets, across *50* randomisations, on Dataset 2. The red line denotes the mean value $r(g,\hat{g})=0.34$ *(s.d. 0.04)* as given in Table 10.

Gene id	Family	Transcript	Chr	Tissue	Function	Disorder	Method	Study
PRDX6L	Peroxiredoxins family	Peroxiredoxin-6-like pseudogene	BTA6	Lung, alveolar macrophages (parental gene)	Intacellular lipid degradation	bTB susceptibility	Heterozygote Disadvantage GWA analysis	Population 1 (Tsairidou et al. Proceedings of the VIth M. bovis conference, 2014)
DHX15	RNA helicase family	RNA helicase (splicing factor)	BTA6	Myeloid dendritic Cells	Viral sensor and innate immune system response (IFN production)	-	RH mapping	Population 1 and Population 2 Meta-analysis (Tsairidou et al. Proceedings of the 10th WCGALP, 2014)
SLC34A2	Solute Carrier family	phosphate transport protein	BTA6	Lung	Phospholipid degradation and phosphate metabolism	Pulmonary Alveolar Microlithiasis (PAM)	RH mapping	Population 1 and Population 2 Meta-analysis (Tsairidou et al. Proceedings of the 10th WCGALP, 2014)
ECSOD (SOD3)	Superoxide Dismutase family	Extracellular Superoxide Dismutase	BTA6	Lung	Antioxidant defence - anti-inflamatory protein	Chronic Obstructive Pulmonary Disease (COPD)	RH mapping	Population 1 and Population 2 Meta-analysis (Tsairidou et al. Proceedings of the 10th WCGALP, 2014)
PTPRT	Protein Tyrosine Phosphatase family	Protein Tyrosine Phosphatase Receptor	BTA13	CNS, liver	Cell growth, differentiation, oncogenic transformation	bTB susceptibility	GWA analysis	Population 1 (Bermingham et al. 2014)
SLC6A6	Solute Carrier family	Taurine Transport protein	BTA22	Macrophages, intestinal epithelial cells	Inflammatory response - host response to bacterial infection	bTB susceptibility	GWA analysis	Population 2 (Finlay et al. 2012)

Table 11. Summary table of candidate genes identified through the Regional Heritability mapping in the present study and genes previously associated with bTB resistance in studies on the populations used in the present meta-analysis.

Appendix 4.1

RH_{max} estimates within each chromosome for the combined Dataset 1 (*n=1438*) for the three window sizes tested. As phenotypes for Population 2 were used the de-regressed EBVs with pedigree information.

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	0.356	0.311	0.091
2	0.316	0.323	0.009
3	0.375	0.346	0.317
4	0.356	0.254	0.022
5	0.299	0.017	0.010
6	0.301	0.322	0.238
7	0.175	0.193	0.014
8	0.319	0.262	0.224
9	0.305	0.294	0.179
10	0.284	0.264	0.013
11	0.286	0.263	0.024
12	**0.409**	0.329	**0.353**
13	0.280	0.241	0.082
14	0.247	0.270	0.019
15	0.255	0.238	0.085
16	0.356	0.318	0.219
17	0.259	0.228	0.188
18	0.322	0.018	0.020
19	0.342	**0.359**	0.007
20	0.393	0.254	0.021
21	0.181	0.220	0.141
22	0.253	0.017	0.010
23	0.362	0.331	0.211
24	0.296	0.045	0.038
25	0.172	0.203	0.010
26	0.289	0.273	0.253
27	0.208	0.177	0.177
28	0.205	0.009	0.008
29	0.285	0.264	0.088

LRT_{MAX} results within each chromosome for the combined Dataset 1 (*n=1438*) and corresponding LRT significance thresholds after the Bonferroni correction, for the three window sizes tested. As phenotypes for Population 2 were used the de-regressed EBVs with pedigree information.

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	16.03	6.88	6.88
2	21.65	17.90	1.08
3	34.82	36.90	**16.52**
4	11.89	4.68	2.52
5	9.57	4.94	2.13
6	26.45	19.22	7.16
7	10.43	4.79	2.93
8	31.11	24.89	6.51
9	38.01	24.81	1.11
10	6.86	5.96	0.79
11	16.41	6.10	5.08
12	16.18	11.76	8.56
13	29.53	4.36	5.51
14	18.67	7.26	5.75
15	25.53	13.13	10.75
16	23.84	17.89	13.54
17	**39.15**	12.31	2.54
18	19.51	5.93	5.14
19	29.96	**47.70**	1.99
20	28.02	12.66	5.54
21	11.26	6.06	1.84
22	12.51	3.25	2.44
23	26.95	15.05	5.51
24	18.92	5.32	3.49
25	2.96	7.14	1.57
26	16.38	17.46	3.53
27	8.85	4.81	4.12
28	12.54	3.38	2.27
29	29.36	32.39	8.04
Genome-wide threshold	16.26	15.50	14.53
Suggestive threshold	10.65	9.90	8.96

Appendix 4.2

LRT results for Population 2 analyses and corresponding LRT significance thresholds, with the pedigree-free EBVs, and the pedigree-free de-regressed EBVs used as phenotypes.

	Window size and step size (number of SNPs)	
Chr	**50, 25 EBVs**	**50, 25 Deregressed EBVs**
1	4.22	3.21
2	9.20	2.67
3	6.58	2.79
4	6.30	**8.74**
5	3.50	7.18
6	2.28	4.78
7	1.48	0.83
8	2.34	3.59
9	0.50	3.62
10	3.70	2.20
11	**10.60**	1.95
12	5.62	2.00
13	3.98	2.70
14	5.28	3.74
15	2.24	7.50
16	2.66	2.88
17	0.58	5.00
18	1.50	0.99
19	3.54	3.26
20	2.72	2.94
21	4.74	7.44
22	6.60	3.59
23	8.72	2.73
24	2.02	1.05
25	1.44	0.89
26	1.84	2.28
27	2.20	1.97
28	2.02	1.13
29	3.94	5.97
Genome-wide threshold	4.73	4.73
Suggestive threshold	9.16	9.16

Appendix 4.3

RH_{max} estimates within each chromosome for Dataset 2, for the three window sizes tested. Pedigree-free EBVs were used for Population 2 after removing the bulls with less than *8* daughters.

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	0.015	0.009	0.012
2	0.008	0.009	0.007
3	0.025	0.017	0.024
4	0.019	0.025	0.019
5	0.015	0.011	0.011
6	0.019	0.023	0.024
7	0.013	0.018	0.015
8	0.014	0.015	0.015
9	0.012	0.010	0.009
10	0.009	0.006	0.006
11	0.015	0.014	0.010
12	0.019	0.017	0.015
13	0.023	0.015	0.024
14	**0.029**	0.021	0.022
15	0.020	0.015	0.017
16	0.017	0.015	0.016
17	0.015	**0.030**	0.036
18	0.010	0.011	0.012
19	0.009	0.009	0.009
20	0.009	0.008	0.013
21	0.012	0.008	0.007
22	0.015	0.016	0.016
23	0.022	0.013	**0.039**
24	0.010	0.008	0.008
25	0.008	0.006	0.009
26	0.011	0.014	0.017
27	0.018	0.022	0.018
28	0.012	0.014	0.012
29	0.025	0.021	0.017

LRT_{MAX} results within each chromosome for Dataset 2 and corresponding LRT significance thresholds after the Bonferroni correction, for the three window sizes tested. Pedigree-free EBVs were used for Population 2 after removing the bulls with less than *8* daughters.

	Window size and step size (number of SNPs)		
Chr	**20, 10**	**30, 15**	**50, 25**
1	3.14	2.45	1.65
2	2.72	1.99	1.64
3	6.04	7.41	6.36
4	4.62	5.53	3.83
5	4.83	2.96	3.51
6	**9.25**	**8.96**	**7.85**
7	3.97	4.48	3.56
8	5.88	5.20	4.38
9	2.02	2.52	1.31
10	1.98	1.60	1.30
11	3.54	3.07	2.64
12	6.70	4.53	3.69
13	2.88	4.22	3.52
14	8.52	7.11	7.33
15	6.40	4.91	3.45
16	4.74	2.66	3.45
17	5.14	4.96	6.88
18	4.48	3.49	2.65
19	1.79	1.65	1.54
20	2.61	1.53	2.13
21	3.01	1.36	0.72
22	7.81	7.21	4.40
23	5.06	2.68	6.88
24	1.67	1.39	0.55
25	1.57	1.55	1.65
26	2.64	2.24	2.74
27	6.43	8.72	5.96
28	4.39	3.95	3.35
29	4.87	4.16	4.02
Genome-wide threshold	16.19	15.43	14.46
Suggestive threshold	10.58	9.83	8.89

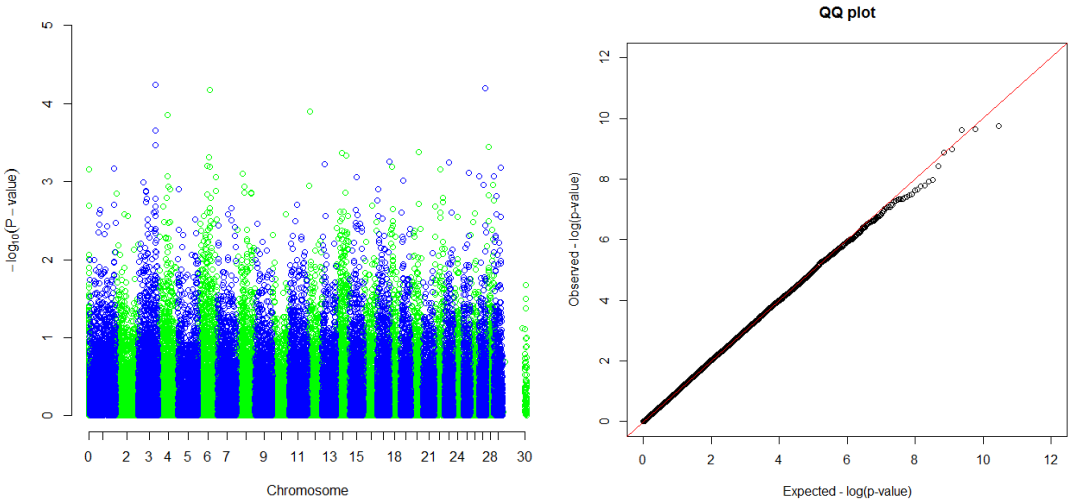
Appendix 4.4

Regional heritability mapping LRT results for the combined analysis of Dataset 2 with zero covariance assumed between the two populations, for the 50-SNP window size (25-SNP step).

	Window size and step size (number of SNPs)
Chr	**50, 25**
1	1.678
2	1.54
3	**8.76**
4	1.98
5	3.81
6	5.95
7	3.11
8	4.32
9	1.74
10	0.30
11	0.95
12	4.30
13	4.44
14	**8.44**
15	4.40
16	2.77
17	3.11
18	5.19
19	2.02
20	1.00
21	0.74
22	6.39
23	1.42
24	1.76
25	1.13
26	1.15
27	5.42
28	2.66
29	2.18
Genome-wide sign threshold	14.46
Suggestive threshold	8.89

Appendix 4.5

GWA analysis Manhattan plot using the *"egscore"* function (GenABEL/R) adjusted for *3* principal components for Dataset 1 and Q-Q plot showing observed compared to expected x^2 values under the null hypothesis of no association. The genome-wide and suggestive significance thresholds were *5.85 and 4.55* respectively.



Appendix 4.6

Chromosomal heritability estimates for Dataset 1, for the four different methods followed: Method (a) where each chromosome was fitted separately ($h_{c(sep)}^2$); Method (b) where all chromosomes were fitted simultaneously (h_c^2); Method (c) where each chromosome was fitted with the whole **G** matrix (h_{c+G}^2); and Method (f) where only the chromosomes with non-zero variance in Methods (a), (b) and (c) were fitted simultaneously ($h_{(non\text{-}zero\ var)}^2$). V_{10} is the sum of the variance explained by the chromosomes whose variance remained non-zero in Method (d).

Chr	Length (Mbp)	$h_{c(sep)}^2$	h_c^2	h_{c+G}^2	$h_{(non\text{-}zero\ var)}^2$	$h_{c(sep)}^2 - h_c^2$
1	161	0.000	0.000	0.000		0.000
2	141	0.004	0.000	0.000		0.004
3	128	0.046	0.042	0.037	0.043	0.004
4	124	0.019	0.003	0.000		0.016
5	126	0.016	0.000	0.000		0.016
6	123	0.054	0.042	0.039	0.043	0.013
7	112	0.020	0.000	0.000		0.020
8	117	0.024	0.018	0.009	0.015	0.006
9	108	0.004	0.000	0.000		0.004
10	106	0.000	0.000	0.000		0.000
11	110	0.010	0.005	0.000		0.004
12	85	0.019	0.000	0.000		0.019
13	84	0.013	0.000	0.000		0.013
14	81	0.061	0.037	0.046	0.042	0.025
15	85	0.030	0.024	0.020	0.020	0.006
16	78	0.008	0.000	0.000		0.008
17	77	0.041	0.015	0.011	0.012	0.026
18	66	0.018	0.002	0.000		0.016
19	65	0.011	0.000	0.000		0.011
20	76	0.016	0.002	0.001	0.000	0.013
21	69	0.001	0.000	0.000		0.001
22	62	0.022	0.020	0.014	0.021	0.002
23	53	0.025	0.010	0.011	0.011	0.015
24	65	0.002	0.000	0.000		0.002
25	44	0.011	0.000	0.000		0.011
26	52	0.006	0.000	0.000		0.006
27	49	0.004	0.004	0.000		0.000
28	46	0.018	0.011	0.008	0.013	0.007
29	52	0.019	0.000	0.003		0.019
V_{10}					0.220	

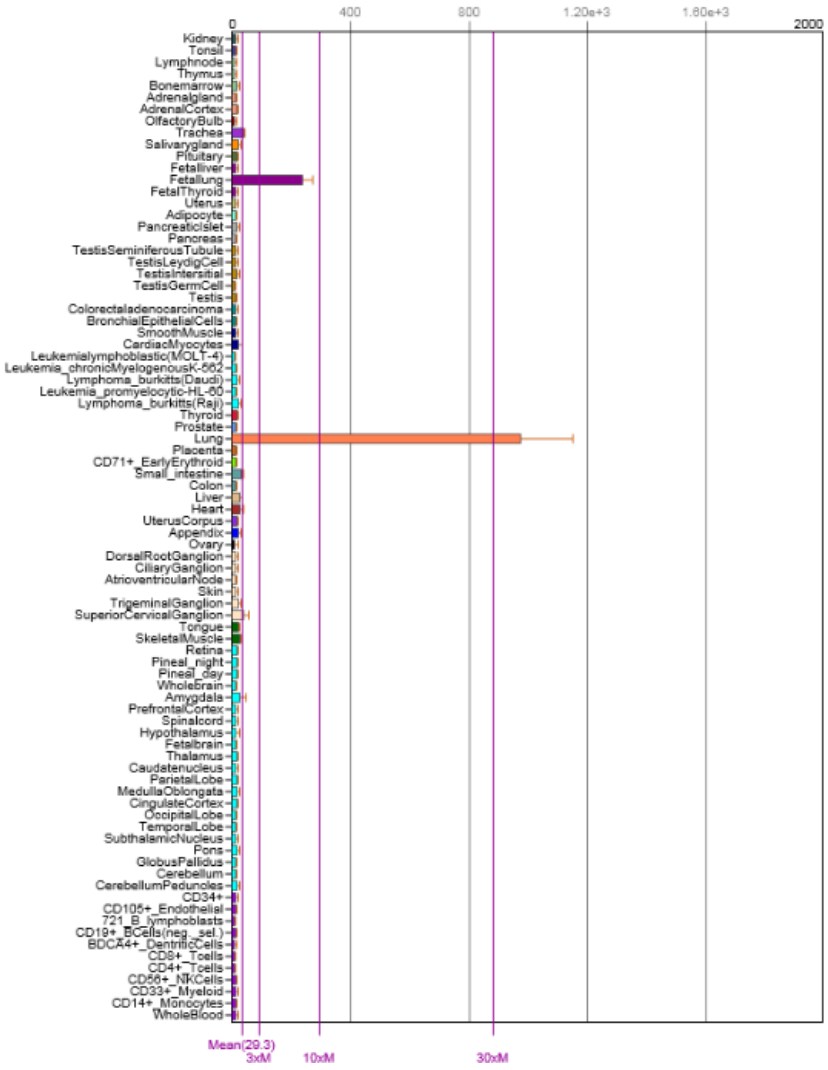
Appendix 4.7

For completeness, in Method (d) for Dataset 2, the chromosomes with non-zero variance in Method (c) were fitted simultaneously plus the $\mathbf{G}_{-c}$ matrix on the remaining chromosomes. In the table are presented the variance explained by each chromosome (V_i) and the corresponding heritability where V_{16} is the sum of the variance explained by the *16* chromosomes fitted simultaneously. In this analysis *16* chromosomes explain *28%* of the total phenotypic variance.

	Method (c)		**Method (d)**	
Chr	$\mathbf{V_i}$	$\mathbf{h^2}$	$\mathbf{V_i}$	$\mathbf{h^2}$
3	0.037	0.036	0.040	0.036
4	0.010	0.010	0.011	0.010
5	0.007	0.007	0.002	0.002
6	**0.048**	**0.048**	**0.059**	**0.053**
7	0.015	0.015	0.012	0.011
8	0.021	0.020	0.023	0.021
14	0.047	0.047	0.044	0.040
15	0.021	0.021	0.014	0.013
16	0.003	0.003	0.004	0.004
17	0.025	0.025	0.034	0.031
22	0.023	0.023	0.022	0.020
23	0.020	0.020	0.021	0.019
25	0.003	0.003	0.000	0.000
26	0.002	0.002	0.003	0.003
28	0.015	0.015	0.020	0.018
29	0.004	0.004	0.000	0.000
$\mathbf{V_{env}}$	-		0.803	
$\mathbf{V_P}$	-		1.115	
$\mathbf{V_{16}}$	0.301		0.312	

Appendix 4.8

Results for mRNA expression for the Solute Carrier family 34 [sodium phosphate], member 2 (SLC34A2). The tissue-specific expression pattern demonstrates that the SLC34A2 candidate gene in the RH mapping analysis is highly expressed in the lung (results taken from biogps gene annotation portal: http://biogps.org/#goto=genereport&id=10568).



Chapter 5

Genotype imputation for dairy cattle: a meta-analysis of directly genotyped and imputed genotypes for bTB resistance

5.1 Introduction

In the previous chapter it was shown that resistance to bovine Tuberculosis is a moderately polygenic, complex trait controlled by a few major chromosomes. Thus, genetic variance for bTB resistance is likely to be explained by several QTLs with moderate individual effects. Furthermore, after having demonstrated that genomic selection for resistance to bTB in cattle is feasible (Tsairidou et al. 2014), both QTL discovery and genomic prediction accuracy are expected to be benefitted by larger datasets with more genotyped and phenotyped animals. Therefore, meta-analysis of different studies can be a powerful method in order to increase the sample size. However, one of the constraints when combining individuals originating from different studies is that they have often been genotyped using different genotyping platforms, and thus, genotypic information is available for different sets of markers. In this context, this chapter expands the analyses presented in the previous chapter through inferring high density genotypes by means of genotype imputation.

More specifically, genotype imputation utilises the presence of haplotypes originating from a distant common ancestor that are shared between apparently unrelated individuals (Identity By Descent (IBD) regions) (Li et al. 2010). The more

distant the common ancestor is the shorter the haplotype stretches will be. Reference haplotypes can either be already available in haplotype libraries (e.g. HapMap CEU) or they can be calculated from the data. To calculate them from the data a pedigree can be used, or alternatively, the haplotypes can be calculated from a reference population genotyped at high density. Then, the "most-likely-genotypes" can be imputed for the un-typed markers in populations genotyped at lower density, by comparison with the reference haplotypes (*"in silico genotyping")* (Burdick et al. 2006; Sanna et al. 2008; Willer et al. 2008; Li et al. 2009; Li et al. 2010). Therefore, genotype imputation provides a useful tool facilitating the combination of different datasets genotyped for a distinct set of SNPs (Willer et al. 2008), allowing us to exploit all the already available information including lower density genotypic data. Further, imputation has been successfully used for fine-mapping of candidate regions by increasing resolution for those regions (Chambers et al. 2008; Sanna et al. 2008; Li et al. 2009), and for improving the power for association analyses (Li et al. 2009).

The aim of this study was to combine in a meta-analysis data containing high density genotypes both directly genotyped and inferred using genotype imputation. The two dairy cattle populations described in the previous chapter were analysed through the means of Regional Heritability Mapping (RHM) and Cross Validation (CV) analysis. We anticipate that these analyses will allow us to identify genomic regions associated with bTB resistance and to investigate the impact of the use of imputed genotypes on the genomic prediction accuracy.

5.2 Materials and Methods

5.2.1 Data description

The dataset comprised merged data from Population 1 comprising Holstein-Friesian cows and including the outliers from PCA, and Population 2 comprising Holstein-Friesian bulls with pedigree-free de-regressed EBVs and including the lower reliability bulls, as described in more detail in the previous chapter (section 4.2.1), however, this time genotypes were those for SNPs on the Illumina BovineHD BeadChip. Before imputation was carried out, Population 1 had been genotyped using the Illumina high-density Bead Chip, while Population 2 had been genotyped with the Illumina Bovine50 SNP chip. The objective was to impute for Population 2 the SNPs that are present on the HD Beadchip but not on the 50K SNP chip as will be described below. Pedigree was not available for either of the populations. Consistent labelling across the reference and the study samples was ensured through obtaining the Illumina forward strand genotypes for both populations. Quality Control (QC) was conducted before imputation with the same quality control criteria being applied to both samples: MAF<0.05, call rate per SNP <95%, HWE $p<0.000001$, completely homozygote and completely heterozygote SNPs were removed, while, unassigned SNPs and SNPs on chromosomes X, and Y were not used for imputation and were excluded from subsequent analysis. The final dataset after QC comprised *573,123* autosomal SNPs for *1,438* individuals (Table 1).

5.2.2 Genotype imputation

Estimation of the haplotypes and genotype imputation of the un-typed markers for Population 2 was conducted for every chromosome using the Markov Chain Haplotyping (MaCH 1.0) package in a four-step process as follows:

5.2.2.1 Selecting the reference panel and calculating the haplotypes

Firstly, the reference haplotypes that will be assumed for the imputation need to be calculated. In the present study the haplotypes were inferred from the data rather than using haplotype libraries derived from other populations. For this purpose, a reference panel of individuals was used, as due to computational limitations it was not possible to use the entire dataset. This panel was selected from the population with the HD genotypes and was selected to be representative of the complete dataset to contain as many as possible of the existing variants and to avoid biasing the imputed genotypes due to breed or case-susceptible/control-resistant differences that may have a genetic background. Therefore, cases and controls in Population 1 were separated in two groups and one animal was selected from each herd within each of the groups. For herds with the greatest contribution to the dataset a second animal was also selected, with this animal being a Friesian where available. This process allowed a selection of *262* animals from *165* herds, from which animals were randomly removed to finally generate a reference panel of *200* animals. This number was chosen as Li et al. (2009) considered that *200* animals would be sufficient for this purpose. The final reference panel selected comprised *160* cases and *40* controls with *164* Holstein and *36* Friesian cows. MaCH software uses a

series of Markov Chain iterations to update the sampled haplotypes and construct a consensus haplotype. The reference haplotypes were calculated from the selected subset after *20* iterations (Li et al. 2010).

5.2.2.2 Estimating model parameters

The cross over map, which determines the likely sites for transition from one haplotype to the next, and the error rate map which flags unusual markers, are model parameters required for the imputation. To make the process computationally efficient, a single set of estimates for the model parameters was obtained using *20* iterations and the *200* reference individuals with the haplotypes as calculated in (5.2.2.1). This, following the recommendations by Li et al. (2010), to avoid using a large number of Markov Chain iterations to simultaneously update the model parameters and impute the missing genotypes.

5.2.2.3 Imputing missing genotypes

These reference haplotypes from Population 1 as calculated in (5.2.2.1) and the model parameters as calculated in (5.2.2.2), were used to infer the most likely genotypes for the un-typed SNPs for Population 2. In the end of this process, HD genotypes were available for all the individuals. The summary statistic r^2 was obtained for each SNP, defined by Li et al. (2010) as the squared correlation between true allele counts and estimated allele counts. This is defined from the number of times each genotype has been sampled after I iterations, where $I = n_{A/A} + n_{A/G} + n_{GG}$, and where e.g $n_{A/G}$ is the number for the heterozygotes. Following Li et al. (2010) a score, e.g. for allele A, was calculated as $g_A = (2\, n_{A/A} + n_{A/G}) / I$). The estimated allele frequency is $p_A = \bar{g}_A / 2$, where $\bar{g}_A$ is the average genotype score over all individuals

for that locus. Then $r^2=var(g_A)/[2p_A(1-p_A)]$, where the denominator is assumed to be the variance of genotype scores if observed without error (Li et al. 2010): it is independent of which allele is used for a biallelic locus.

5.2.2.4 Assessment of imputation quality

In order to assess the overall performance of the imputation process, a proportion of genotypes was randomly masked in MaCH, allowing comparison of the imputed genotypes at these locations with the true genotypes. For this purpose, *2%* of the reference genotypes were masked for each chromosome (http://csg.sph.umich.edu/abecasis/MACH/tour/), where the number of masked genotypes was equal to (the number of SNPs per chromosome) x (the number of individuals in the reference population) x *0.02*. The r^2 values defined as the squared correlations between the imputed genotypes and the true genotypes, the allelic error rates after masking, the estimated per genotype error rate, the estimated per allele error rate, and the estimated mismatch rate in the Markov model were obtained after *20* rounds.

Furthermore, in order to investigate the impact of the quality of the imputed genotypes on heritability estimates and on genomic prediction accuracy, the r^2 calculated in (5.2.2.3) was used to identify lower quality imputed genotypes and discard them from subsequent analyses. This measure provides an estimate of the between replicate variability, and can be used to identify markers that give consistently poor values. Two different filters were applied, by setting as "missing" the imputed genotypes corresponding to markers having (a) an $r^2 < 0.7$ (Filter 1) and

(b) an $r^2 < 0.9$ (Filter 2), and consistency of results across the different thresholds was examined.

5.2.3 Data analysis

The combined data, containing all the directly genotyped SNPs and the filtered imputed genotypes, were analysed following the regional heritability mapping and the cross validation methodologies in order to identify genomic regions associated with bTB resistance and investigate the impact of using imputed genotypes on genomic prediction accuracy.

5.2.3.1 Regional heritability mapping

Regional heritability mapping methodology was followed as described in the previous chapter (section 4.2.2.1) with the difference that for this dataset the selected size of the overlapping windows was larger (i.e. 200 SNPs, with a 100-SNP step) to account for the denser SNP chip. This method allows us to obtain heritability estimates for genomic regions and identify regions associated with the trait of interest through the Likelihood Ratio Test (LRT). In order to investigate the impact of the imputation quality on the heritability estimates, RH analyses were conducted using (a) all the imputed genotypes, (b) filtered imputed genotypes for an r^2 threshold of *0.7*, and (c) filtered imputed genotypes for an r^2 threshold of *0.9*.

5.2.3.2 Genomic prediction

The accuracy of the genomic prediction calculated as $E[r(g,\hat{g})] \approx r(y,\hat{y})/h$, using both directly genotyped and imputed genotypes was estimated through a 5-fold cross validation (CV) (Luan et al. 2009) as described in the previous chapter (section

4.2.2.4). Individuals from both populations were combined and randomly assigned to five groups of near-equal size. This process was repeated *50* times to obtain the average accuracy across *50* different randomisations of the individuals into groups and examine the variability across the repeats.

To investigate the impact of the quality of the imputed genotypes on the prediction accuracy this analysis was repeated using (a) all the imputed genotypes, (b) filtered imputed genotypes using Filter 1, (c) filtered imputed genotypes using Filter 2, and (d) imputed genotypes only for SNPs on two chromosomes found to contribute to the observed variation (Chapter 4 section 4.3.2 and 4.3.4).

5.2.3.3 Targeted imputation

As it was demonstrated in the previous chapter, loci which do affect resistance are spread across a number of chromosomes but, critically, not all chromosomes, suggesting that there are a few major chromosomes affecting the trait. Therefore, prediction accuracy might benefit by targeted imputation for the chromosomes explaining most of the genetic variation. To test this hypothesis cross validated prediction accuracy was estimated using imputed genotypes only for the SNPs on BTA6 (after applying Filter 2), while for the rest of the chromosomes were used the overlapping genotypes between the low density and the HD SNP chips. BTA6 was selected for the following reasons: (a) there was indication that BTA6 might be associated with the trait in the RH mapping in 5.2.3.1 using the imputed genotypes, (b) the RH mapping analysis presented in the previous chapter revealed an association of a region on chromosome 6 with bTB resistance at the suggestive level (section 4.3.2), and (c) the chromosomal heritability estimation analysis

presented in Chapter 4 (section 4.3.4) has shown that chromosome 6 is explaining a large proportion of the observed variation.

5.3 Results

5.3.1 Imputation

Table 2 shows the r^2 obtained after masking *2%* of the genotypes as a measure of imputation quality (section 5.2.2.4 (Table 2). The average r^2 after masking across all chromosomes was *0.979,* ranging from *0.974* (BTA26) to *0.983* (BTA8), indicating an overall good quality of the imputed genotypes (Fig. 1). In all cases the maximum r^2 across all SNPs was *1*. Further details about the masking process including the percentage of markers masked and the time needed for the process to be completed are presented in Table 3. The mean allelic error rate after masking was *0.0013*, further indicating a very good imputation quality.

The average r^2 within each chromosome without masking as calculated in 5.2.2.3, was found to range from *0.976* (BTA26) to *0.984* (BTA8), with a mean of *0.981* across all chromosomes. The minimum r^2 value was observed on BTA3 (r^2*=0.517*) (Table 2, Fig. 2). *0.10%* of the SNPs was excluded from the data using Filter 1 (r^2*<0.7*), and *0.76%* using Filter 2 (r^2*<0.9*) which resulted in a marginal increase in the average r^2 to *0.982* (Table 2). For individual chromosomes the impact on the r^2 depends on the number of SNPs filtered. These r^2 values estimated from the genotype scores were slightly higher compared to those after masking, however they were in good agreement for all the chromosomes. Unsurprisingly, the r^2*<0.9* threshold provided better r^2 compared to the r^2*<0.7* threshold, but also it provided a

better estimate compared to the r^2 after masking *2%* of genotypes (Table 2).

5.3.2 Genomic heritability estimates

ASReml analysis on the combined imputed data provided a genomic heritability estimate of $h^2 = 0.129$ *(SE = 0.048),* while after discarding the imputed genotypes with $r^2<0.7$ or $r^2<0.9$, genomic heritability remained practically unchanged with estimates of $h^2 = 0.127$ *(SE = 0.048)* and $h^2 = 0.126$ *(SE = 0.048)* respectively (Table 4). This estimate is similar to the heritability obtained in the previous chapter (*0.14 (SE = 0.05)*) when combining Population 1 and Population 2 and using only the SNPs present in common in the HD and low density SNP chips, while it is lower, although within the SE difference, than the h^2 of *0.19 (SE = ±0.06)* obtained after removing the lower reliability bulls (section 4.3.1).

5.3.3 Regional heritability estimates

Regional heritability estimates are presented in Table 5 and applying the filters was not found to have a large impact on the estimates (Table 5). RH mapping using all the imputed genotypes did not reveal any significant associations according to LRT (Table 6). However, after applying the $r^2<0.7$ and the $r^2<0.9$ filters, there was an association of a region on BTA 8, significant at the suggestive level (Fig. 3). After applying the filters, BTA6 was also very close to the suggestive significance threshold (Fig. 3b and c). For BTA6, RH mapping using imputed genotypes confirmed the regions which were identified in Chapter 4 when using only the overlapping HD and low density SNPs: (a) the first most significant region on BTA6 after imputation is contained within the second most significant region before

imputation, and (b) the second most significant region after imputation partly overlaps with the first most significant region before imputation (Table 7).

5.3.4 Genomic prediction

Cross validation when all the imputed genotypes were included in the data provided an average prediction accuracy of $E[r(g,\hat{g})] = 0.323$ *(s.d. 0.058)* (Fig. 5a). After discarding the SNPs with $r^2 < 0.7$ and $r^2 < 0.9$ the prediction accuracy estimates were *0.317 (s.d. 0.052)* and *0.332 (s.d. 0.046)* respectively. Variability across the cross validation repeats was reduced when the most stringent filter was applied compared to when using all the imputed genotypes (Fig. 5c).

5.3.5 Targeted imputation

Following the targeted imputation approach for BTA6, i.e. using imputed genotypes for BTA6, while low density genotypes were used for the rest of the genome, a prediction accuracy of *0.410 (s.d. 0.052)* was obtained, and the variability across the CV repeats was reduced with no replicate providing an accuracy less than *0.3* (Figure 6a). This improvement was not due to changes in the estimate of the heritability as the correlation between the predicted breeding value and the observed phenotype also improved compared to genome-wide imputation *(0.13 vs 0.11)*.

Using targeted imputation for BTA8 provided a prediction accuracy of *0.379 (s.d. 0.052)*. However, this analysis provided a lower correlation *(0.11)* and some CV repeats provided accuracies lower than *0.3* (Figure 6b). Lastly, imputed genotypes were used simultaneously for both BTA6 and BTA8, while low density genotypes

were used for the rest of the genome. This analysis provided an accuracy of *0.407 (s.d. 0.054)* (Table 6).

Further, BTA13 has been previously associated with bTB resistance in a study on Population 1 (Bermingham et al. 2014), however, there was no evidence for association for this chromosome in this data after combining Population 1 and Population 2. The targeted imputation approach as described above was tested for BTA13 and provided reduced accuracy *(0.341 (s.d. 0.086))* and correlation with phenotypes *(0.08)* compared to BTA6, while the results across the repeats were highly variable (Fig. 6c).

5.4 Discussion

5.4.1 Overall success of genotype imputation for cattle data

Genotype imputation for dairy cattle data was found to be successful when imputing from the Illumina Bovine SNP50 Bead Chip to the Illumina Bovine HD Bead Chip, with a mean r^2 across all chromosomes of *0.98*. Genotype imputation allowed us to exploit all the available information from the low density dataset in a cost-effective way, as high density, direct genotyping is required only for a small subset of animals which were used as the reference population.

Each genotype generated through imputation by MaCH is a prediction of the "most likely estimate", and therefore it is calculated with some uncertainty. This uncertainty is introduced by several factors including the quality of markers used as the reference for the calculation of the haplotypes and the genetic distance between the reference population and the sample-to-be-imputed. Recombination and mutation

events will disrupt the haplotypes and the more distant the populations are, the stronger this effect would be. In the present study this is pertinent because high density genotypes were available only for one population and imputation was conducted across different populations. However, both populations comprised Holstein and Friesian cows and these breeds are not expected to be very distantly related (http://www.ukcows.com/holsteinuk/publicweb/Services/SrvMain.aspx?page=_BriefHistory&cmh=166). Moreover, Li et al. (2009) demonstrated that imputation accuracy increases with reference populations of larger size, with little benefit being observed when further increasing the reference panel beyond *200* individuals. Therefore, in this data, *200* individuals *(~14%* of the data) were used as a reference panel. Lastly, one of the advantages of imputation studies in livestock is the use of pedigrees so that long haplotypes can be tracked in the population and used in the imputation. However, in the present study pedigree was not available and haplotypes were inferred utilising short haplotype stretches inherited from an unknown distant common ancestor and shared among apparently unrelated individuals, which is the approach commonly adopted for imputation in humans.

All the factors described above are likely to have a negative impact on the imputation accuracy. However, despite those reservations, the apparent accuracy of imputation in this study was high. Previously reported imputation accuracy has been generally high for cattle. Similarly to our findings, Hoze et al. (2013) have reported an imputation accuracy *>97%* across 16 different cattle breeds, when using the Beagle software to impute from the Bovine SNP50 to BovineHD BeadChip, without taking into account pedigree information for the haplotyping. While Binsbergen et al. (2013), report lower accuracies of *0.77-0.83,* depending on the reference population

size, when imputing to whole-genome sequence data. However, caution should be taken in comparing the imputation accuracy values across different studies as the methods for the accuracy calculation vary depending on the software, the selected imputation quality measure and the method of calculation. Although not directly comparable, these values are indicative of an expectation of overall high accuracy values for genotype imputation in cattle.

The imputation software MaCH has been previously used in human studies for the identification of loci associated with variation in height, by means of a GWAS in a meta-analysis of directly genotyped and imputed genotypes for individuals originating from different studies, genotyped using different genotyping platforms (Sanna et al. 2008). Genotype imputation was used to facilitate comparison between different studies and association analyses of the combined data. Further, in a meta-analysis by Willer et al. (2008), a GWAS on combined populations through the use of MaCH, identified new loci affecting the risk of coronary artery disease. Human height and coronary artery disease risk are complex traits controlled by a large number of variants, and genotype imputation was useful in validating or identifying new QTLs. However, the imputation in those studies was based on relatedness information and HD individuals were available in the same population as the individuals-to-be-imputed which were offspring and siblings of the HD individuals. Moreover, for humans there are available datasets of reference haplotypes (International HapMap CEU Project), while in the present study for bTB resistance the reference haplotypes were estimated from the data.

This study used a much higher threshold for filtering compared to what is commonly used in the literature. A threshold suggested for quality assessment of the

imputed genotypes is that of $r^2>0.3$ (Sanna et al. 2008; Li Yun et al. 2010). However, in the present study the minimum r^2 value observed was $r^2_{min}=0.549$ (on BTA3) and therefore the thresholds of $r^2\geq0.7$ and $r^2\geq0.9$ were used. The density of SNPs required depends upon the species and the breed i.e. its effective population size, the range of the LD, the specific genetic architecture of the trait under study, and the trait heritability (Bishop et al. 2010, Chapter 1). Thus, a possible explanation why genotype imputation for cattle was found to be more successful compared to human studies, is that in Holstein cattle, the smaller effective population size ($Ne \approx 50\text{-}100$) and therefore, the longer range LD are expected to allow for longer haplotypes. Thus, relatively fewer markers should be required to achieve a reasonably good imputation quality. Additionally, the density of 50K SNP chip used for genotyping is near the plateau of the curve relating marker density to accuracy (Hayes et al. 2010) and thus it is perhaps unsurprising that that it was sufficient to deliver good imputation accuracy.

Specifically for the chromosomes identified in the RH mapping analysis, the mean r^2 after masking was very close to *1* ($r^2=0.981$) indicating a good imputation quality for those chromosomes (Fig. 3 and Fig. 4). However, the masking is likely to be an overestimate since the reference haplotypes have been calculated using the same reference population. Therefore, for further assessing the imputation quality and the model parameters for BTA6, a subset of *200* HD individuals was randomly selected (excluding the reference pop to produce an independent validation set) and reduced to low density. Then using the haplotypes as calculated from the reference panel, and the same error rate and cross over maps as calculated in 5.2.2.2, those individuals were imputed back to HD and new r^2 correlations were obtained

between the imputed and the observed genotypes for those *200* individuals. This approach provided a more objective correlation of *0.955*, which is lower than that obtained after masking, nevertheless, it still indicates a very good imputation quality for this chromosome. This estimate is closer to the estimates reported by Hoze et al. (2013) and Binsbergen et al. (2013), although it is unclear whether an independent validation set had been used.

5.4.2 Genotype imputation and heritability estimation

One of the hypotheses was that higher heritabilities would be obtained by using imputed genotypes. However, using imputed genotypes was not found to have a large impact neither on the genomic heritability or on the regional heritability estimates. Genomic heritability obtained in this chapter was very similar to the estimate obtained when using only the overlapping SNPs between the HD and low density SNP chips *(h^2=0.14 (SE = 0.05))* (Chapter 4, section 4.3.1).

The second hypothesis was that using imputed genotypes might allow detecting better association signals. In previous studies genotype imputation has either assisted in identifying new independent variants or enhanced formerly identified associations (Burdick et al. 2006; Willer et al. 2008). In the present study RH mapping using imputed genotypes (a) identified a region on BTA8 significant at the suggestive level which was not detected when using only the overlapping HD and low density SNPs (Chapter 4), and (b) confirmed the previously identified associations on BTA6. In conclusion, although imputation can be useful to enhance signals, using stringent quality filters has an important effect on the power to detect associations.

5.4.3 Genotype imputation and genomic prediction

The motivation for this analysis was the same as in Chapter 4 i.e. prediction accuracy is expected to be improved by increasing the sample size. While in Chapter 4 increased sample size was achieved by combining two populations and reducing the genotypes to the lowest density to use the overlapping SNPs between the HD and the 50K SNP chips, in this chapter, HD genotypes were obtained by means of genotype imputation. Greater SNP density would be expected to improve the prediction accuracy by providing more information on the relationships. However, the set of additional imputed animals was relatively small in this study and arguably it was not large enough to have a detectable impact on the prediction accuracy. Further, each genotype is a "most likely estimate", estimated with some uncertainty which builds additional errors in the **G** matrix. Consequently, accuracy was not improved when using all the imputed genotypes.

Filtering out markers based on their r^2 did not have a significant impact on the genomic prediction accuracy, which can be explained by the fact that since bTB resistance is not expected to be controlled by one major gene and control is spread over several genes across the genome, even if imputation quality might not be great for some SNPs within some of those genes, their effect will be small. The prediction accuracy estimates obtained in this analysis are similar to the estimates obtained when using only the overlapping SNPs between HD and low density SNP chips (i.e. *0.33 (±0.05)*). The regression of the number of SNPs per chromosome when only the overlapping SNPs are used, and when imputed HD markers are used, shows that the relative weight of the chromosomes to the total number of SNPs has not changed and

thus these analyses are directly comparable (Fig. 7).

These results demonstrate that the limiting factor is not the marker density, but the number of animals and the trait definitions. This result is consistent with similar findings from previous studies. Sánchez-Molano et al. (2015) observed for a UK Labrador retriever population, that the GBLUP prediction accuracy when estimated for different numbers of randomly selected markers, did not further improve when increasing the number of SNPs above 11K SNPs. A possible explanation is that genomic prediction accuracy depends on the genetic structure of the species under study i.e. the number of independent chromosome segments which depends on the effective population size (Chapter 2, section 2.2.10.2) (Daetwyler et al. 2010). Thus, the long-range LD in dogs and in cattle, and the smaller effective population sizes, are likely to provide improved prediction accuracies for lower density SNP chips i.e. for the same statistical power, the SNP density required is reduced (Daetwyler et al. 2010; Quilez et al. 2012). Other studies in livestock have also reported diminishing benefits in GBLUP prediction accuracy by increasing SNP density beyond a certain point. Hayes et al. (2010) demonstrated that genomic prediction accuracy depends on the genetic architecture of the trait, and when the trait is controlled by QTLs with large effects, a very small number of SNPs in LD with those QTLs would be required to reach a sufficient accuracy, with very little benefit when increasing further the number of SNPs. For a commercial population of broiler chickens, Ilska et al. (2014) found that increasing the chip density above the 19K density generated a negligibly small increase in the accuracy of GEBV prediction.

5.4.4 Genomic prediction with targeted imputation

BTB resistance was shown in the previous chapter to be a moderately polygenic trait with some chromosomes estimated to contribute zero variance and a few, major chromosomes contributing considerably more than average. Genomic prediction accuracy may be benefitted by taking into account the specific genetic architecture of the trait under study. There is a previous example (Serão et al. 2014) of a disease resistance trait where region specific analysis taking into account the specific genetic architecture of the trait provided more accurate genomic prediction and identified associations that were not visible when analysing genome wide and looking at all chromosomes simultaneously. Antibody response to PRRS in pigs was found to be largely controlled by two QTLs on SSC7. Serão et al. have reported improved prediction accuracy for response to PRRS when only SNPs residing in those QTLs were analysed (prediction accuracy of 0.63 when analysing only the SNPs on SSC7, versus 0.49 for all the SNPs across the genome) (Serão et al. 2014). This indicates that when the trait under study is controlled by a few QTLs with relatively large effects, adding to the analysis the remaining chromosomes that have no effect may add noise causing dilution of these effects. However, bTB is an intermediate situation where there is not a major QTL but genetic variance is clustered over only a handful of QTLs with moderate effects, and thus distinguishing those effects from noise is more challenging.

Prediction accuracy was found to be improved when using HD imputed genotypes only for the chromosomes for which there was prior indication that it is associated with the trait while overlapping low density genotypes were used for the

rest of the genome, combined into a single **G** matrix. The effect of this approach was to increase the resolution adding dense genotypes locally, and provided higher prediction accuracy and less variability across the cross validation repeats. These putative QTLs have been identified in the same population and thus, the results may be only reinforcing some cryptic structure within the data. Therefore, in order to validate the QTLs, this finding should be confirmed in another independent population. Consequently, targeted imputation may be very effective where there is independent prior information but should be treated with caution otherwise. However, under the assumption that it is a true QTL, in agreement with what was shown by Serão et al., this result suggests that accuracy might benefit from adding weight on the QTLs associated with the trait under study (Serão et al. 2014). In the present analysis, additional weight on the chromosome containing the QTL was placed in the calculation of the IBS matrix by including HD genotypes only for that chromosome while low density genotypes were used for the rest of the genome.

An alternative approach would be to calculate two separate IBS matrices, one for the HD genotypes for BTA6, and one for the rest of the genome excluding the chromosome of interest i.e. $y = \beta + g_c + g_{-c} + e$. This approach avoids placing additional weight to the putatively associated chromosome in the calculation of a single IBS matrix and adds insight into the source of the benefit for the prediction accuracy. For example, it may be the genetic architecture of the trait or the additional information provided by the imputed genotypes. This approach provided a prediction accuracy of *0.368 (s.d. 0.05)*, which was greater than the accuracy obtained when not considering BTA6 separately. To test if this value was significantly different than when not using any HD genotypes, two IBS matrices were again calculated

separately for BTA6 and for the rest of the genome, but without using the imputed genotypes for BTA6. This analysis provided a prediction accuracy of *0.361 (s.d. 0.04)*. In both cases there was a small increase in the prediction accuracy of the order of a standard deviation. The similarity of the two estimates indicates that the observed increase is not due to the use of HD genotypes. Moreover, these accuracy estimates are not directly comparable to the estimates described earlier, due to differences in the calculation of the heritability and thus the accuracies. For this analysis, the accuracy is calculated as the correlation between the phenotypes, divided by the square root of the sum of the heritabilities as calculated in ASReml. However, in agreement to what was discussed above, even with a putative QTL on BTA6, it was not possible to demonstrate in this data a significant improvement by using denser SNPs, leading to the conclusion that the limiting factor is not the density of the markers.

Lastly, genomic prediction using directly genotyped and imputed genotypes only for BTA6 and ignoring the rest of the chromosomes resulted in an average prediction accuracy of *0.55 (s.d. 0.08)*. When this analysis was repeated using only the low density genotypes for BTA6, an accuracy of *0.52 (s.d. 0.08)* was obtained. Thus, the increase in accuracy could not be shown to originate from the use of imputed genotypes but from using BTA6 alone, on which RH mapping and chromosomal heritability estimation analyses in the same data have indicated that there are putative QTLs. This is the same phenomenon as observed by Serão et al. (2014), and whilst it is a striking increase of accuracy which is hard to ignore, validation in a different population would be necessary as this result is likely to be strongly linked to the features of the particular dataset under study. Further, this

approach ignores the genetic variance explained by other chromosomes except through the similarity of genomic relationships of BTA6 to the remaining chromosomes, i.e. arising from family structure. Given the polygenic nature of the trait and the results from the chromosomal heritability estimation analyses (see previous chapter), most likely there are more QTLs on other chromosomes.

5.4.5 Conclusion

Genotype imputation was found to be successful for dairy cattle data, but using imputed genotypes genome-wide added little information given the sample size and for the trait under study. These findings suggest that the limiting factor is the number of animals and the phenotype definitions rather than the density of genotypes. As it was demonstrated in the previous chapter, loci which do affect resistance are spread across a number of chromosomes but, critically, not all chromosomes. With bTB resistance being a moderately polygenic trait controlled by a few major chromosomes, genomic prediction accuracy might be improved when taking into account this specific genetic architecture of the trait. The approach of targeted imputation as presented in this chapter may be beneficial, but the conclusion would rely on independent validation of the suggestive QTLs documented in earlier chapters.

	n animals	n SNPs after QC
Population 1	1151	**588332**
Population 2	287	**41418**
Combined Dataset	1438	**588332 (573123 autosomal)**

Table 1. Number of animals and number of SNPs after Quality Control (QC), for each population and for the combined imputed dataset.

	Before filtering		(a) $r^2 \geq 0.7$		(b) $r^2 \geq 0.9$		(c) After masking	
Chr	r^2_{mean}	r^2_{min}	**% of SNPs filtered out**	r^2_{mean}	**% of SNPs filtered out**	r^2_{mean}	r^2_{mean}	r^2_{min}
1	0.983	0.63	0.16	0.984	1.33	0.985	0.981	0.632
2	0.981	0.607	0.13	0.982	0.53	0.982	0.979	0.614
3	0.982	0.517	0.09	0.982	0.68	0.983	0.98	0.549
4	0.981	0.586	0.06	0.981	0.70	0.982	0.979	0.61
5	0.981	0.602	0.08	0.981	0.54	0.982	0.979	0.618
6	0.983	0.597	0.10	0.984	1.18	0.985	0.981	0.61
7	0.981	0.565	0.21	0.982	0.76	0.983	0.979	0.59
8	0.984	0.659	0.03	0.985	0.44	0.985	0.983	0.656
9	0.978	0.606	0.23	0.979	0.68	0.98	0.976	0.617
10	0.979	0.662	0.12	0.98	1.39	0.982	0.978	0.656
11	0.982	0.571	0.07	0.982	0.77	0.983	0.98	0.556
12	0.98	0.702	0.00	0.98	0.50	0.981	0.977	0.709
13	0.984	0.611	0.16	0.984	0.39	0.984	0.983	0.648
14	0.983	0.552	0.03	0.984	0.45	0.984	0.983	0.571
15	0.98	0.604	0.41	0.982	1.47	0.983	0.978	0.608
16	0.979	0.645	0.08	0.979	1.52	0.981	0.976	0.665
17	0.982	0.604	0.02	0.982	0.45	0.983	0.981	0.621
18	0.981	0.691	0.01	0.981	1.13	0.982	0.979	0.708
19	0.98	0.654	0.01	0.98	0.22	0.98	0.978	0.676
20	0.982	0.595	0.06	0.982	0.34	0.982	0.98	0.599
21	0.981	0.561	0.08	0.981	1.08	0.983	0.979	0.576
22	0.979	0.606	0.13	0.98	0.61	0.98	0.977	0.615
23	0.982	0.556	0.33	0.983	0.51	0.983	0.98	0.587
24	0.98	0.633	0.06	0.98	0.66	0.981	0.979	0.66
25	0.979	0.831	0.00	0.979	0.38	0.98	0.978	0.835
26	0.976	0.685	0.02	0.976	1.09	0.978	0.974	0.687
27	0.978	0.703	0.00	0.978	0.39	0.979	0.976	0.703
28	0.981	0.739	0.00	0.981	0.08	0.981	0.979	0.743
29	0.98	0.638	0.08	0.98	0.26	0.98	0.979	0.658
mean	0.981	0.628		0.981		**0.982**	0.979	0.64

Table 2. Mean r^2 for every chromosome with the minimum r^2 that was observed within the chromosome, as resulted from the imputation process before and after filtering out the markers with (a) $r^2<0.7$ and (b) $r^2<0.9$, and (c) after masking *2%* of the genotypes.

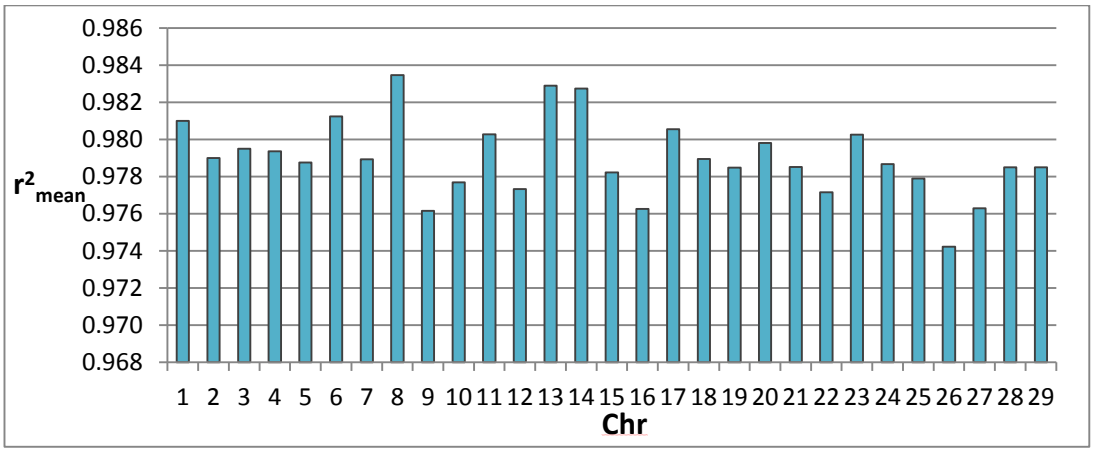


Figure 1. Mean imputation accuracy *(r^2)* for every chromosome estimated as the squared correlation between the imputed genotypes and the true genotypes, after *2%* masking.

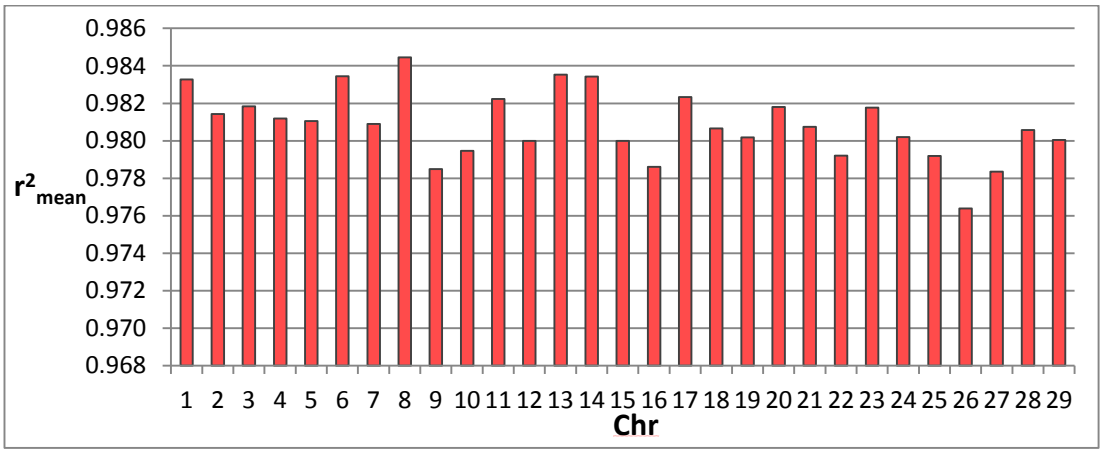


Figure 2. Mean imputation accuracy *(r^2)* for every chromosome as estimated from the genotype scores.

Chr	Number of SNPs per chr	Number of masked Genotypes	Genotype error rate *10^{-3}	Allelic error rate *10^{-3}	Mismatch rate *10-3	Time (h)
1	36196	158376	2.1	1.1	0.9	36.8
2	30659	133178	2.3	1.2	0.8	29.9
3	27653	120396	2	1	0.9	27.3
4	27491	119981	2.6	1.3	0.8	28.1
5	26592	114875	2	1	0.9	25.6
6	28539	124516	2.1	1.1	0.8	28.5
7	25338	110314	2.6	1.3	0.8	24.6
8	21415	94514	2.3	1.2	0.8	21.3
9	24358	105366	2.4	1.2	0.9	24.0
10	24489	106349	2.1	1.1	0.9	24.0
11	25956	112716	2	1	0.9	25.2
12	20493	88629	2.1	1.1	1	19.8
13	15882	70489	3	1.5	1	15.7
14	16290	72550	2.9	1.5	1	16.3
15	19782	85600	2.3	1.2	0.9	19.2
16	18880	81605	2.1	1.1	1	18.1
17	18055	78732	2.5	1.3	0.9	18.5
18	16077	69493	2.3	1.2	0.9	15.8
19	15402	66756	2.3	1.2	1	15.1
20	17725	77206	2.5	1.3	1	17.7
21	16416	71050	2.2	1.1	0.9	15.7
22	15098	65344	3	1.6	1	14.6
23	12477	54021	1.9	1	1.2	12.3
24	14588	63308	2.6	1.3	1	14.3
25	10817	47062	3.1	1.6	1.1	10.5
26	12598	54605	2.6	1.4	1.1	12.8
27	10991	47835	3.2	1.7	0.9	11.4
28	11007	47811	3	1.5	1.1	10.7
29	11859	51504	3	1.5	1	11.5
Mean	**19763**	**86006**	**2**	**1.3**	**1**	**19.5**

Table 3. Descriptive values of the masking process with information on the number of markers masked, the estimated per genotype error rate, the estimated per allele error rate, the estimated mismatch rate in the Markov model, and the time needed for the process to be completed, for every chromosome.

	h^2_{gen}	SE
Population 1	0.23	0.06
Population 2	0.00	0.22
Overlapping SNPs	0.14	0.05
All SNPs with imputation	**0.1286**	0.0479
Filter 1 (r^2≥0.7)	0.1267	0.0476
Filter 2 (r^2≥0.9)	0.1264	**0.0475**

Table 4. Genomic heritability estimates for the two populations and for the combined dataset.

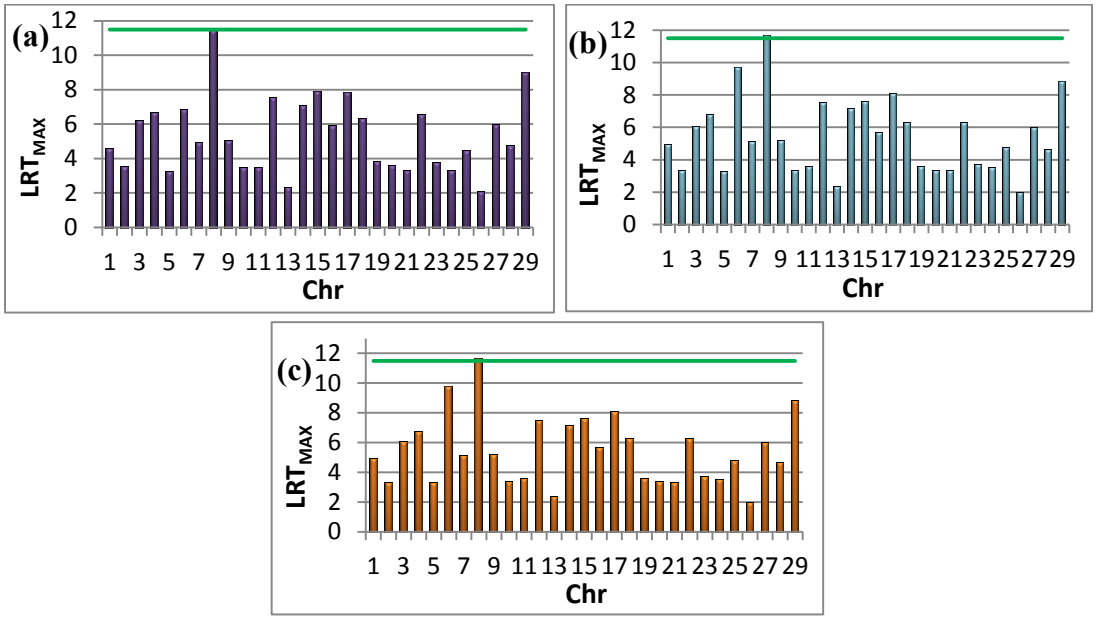


Figure 3. LRT_{MAX} values identified by the RH mapping for every chromosome (a) when using all imputed genotypes, (b) after applying Filter 1 hence removing all SNPs with $r^2<0.7$, and (c) after applying Filter 2 hence removing all SNPs with $r^2<0.9$. The horizontal green lines represent the suggestive threshold.

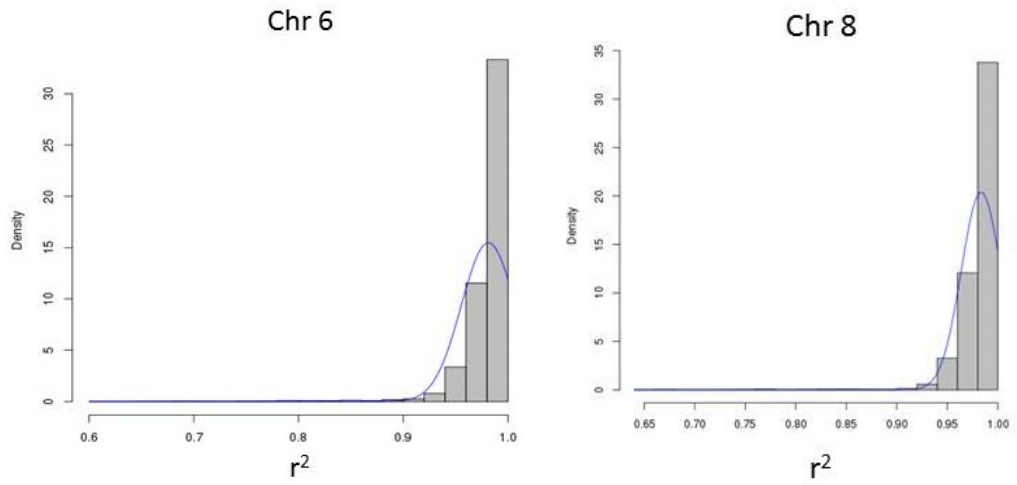


Figure 4. r^2 histograms after masking *2%* of the genotypes for BTA6 and BTA8 identified in the RH mapping analysis on the combined imputed data. Values close to *1* indicate good imputation quality.

Chr	All imputed (200, 100)	Filter 1 (200, 100)	Filter 2 (200, 100)
1	0.022	0.023	0.023
2	0.008	0.008	0.008
3	0.024	0.025	0.025
4	0.026	0.027	0.027
5	0.012	0.012	0.012
6	0.016	0.035	**0.089**
7	0.012	0.012	0.012
8	0.018	0.019	0.019
9	0.010	0.010	0.010
10	0.012	0.012	0.012
11	0.020	0.020	0.020
12	0.023	0.022	0.022
13	0.008	0.009	0.009
14	**0.050**	**0.051**	0.051
15	0.016	0.015	0.015
16	0.015	0.015	0.015
17	0.026	0.023	0.023
18	0.013	0.013	0.013
19	0.013	0.013	0.013
20	0.007	0.007	0.007
21	0.009	0.009	0.009
22	0.019	0.019	0.019
23	0.014	0.015	0.015
24	0.016	0.016	0.016
25	0.011	0.011	0.011
26	0.011	0.011	0.011
27	0.018	0.018	0.018
28	0.017	0.017	0.017
29	0.028	0.027	0.027

Table 5. RH_{max} estimates within each chromosome for the combined dataset from the RH mapping analysis (200-SNP window size and 100-SNP step size) when (a) using all the imputed genotypes, (b) after applying Filter 1, and (c) after applying Filter 2. For each window size the maximum value is in bold.

Chr	All imputed (200, 100)	Filter 1 (200, 100)	Filter 2 (200, 100)
1	4.510	4.886	4.876
2	3.478	3.280	3.280
3	6.154	6.026	6.036
4	6.640	6.726	6.730
5	3.210	3.246	3.250
6	6.796	9.678	9.690
7	4.880	5.056	5.060
8	**11.386**	**11.594**	**11.598**
9	4.984	5.118	5.122
10	3.438	3.314	3.316
11	3.444	3.558	3.546
12	7.490	7.476	7.464
13	2.276	2.278	2.278
14	7.024	7.092	7.084
15	7.850	7.542	7.550
16	5.882	5.614	5.614
17	7.824	8.048	8.042
18	6.266	6.244	6.242
19	3.770	3.546	3.554
20	3.584	3.296	3.292
21	3.262	3.274	3.282
22	6.498	6.240	6.252
23	3.708	3.670	3.672
24	3.278	3.452	3.446
25	4.432	4.698	4.720
26	2.024	1.918	1.924
27	5.904	5.922	5.920
28	4.692	4.598	4.610
29	8.960	8.762	8.770
Genome-wide threshold	17.131	17.131	17.131
Suggestive threshold	11.495	11.495	11.495

Table 6. LRT_{MAX} results within each chromosome with 200-SNP window size and 100-SNP step size, for the combined dataset when (a) using all the imputed genotypes, (b) after applying Filter 1, and (c) after applying Filter 2, and LRT significance thresholds after the Bonferroni correction.

RH mapping		Chr	Start SNP position (bp)	End SNP position (bp)
(1) All imputed genotypes	**(a)**	6	72,915,853	73,645,837
	(b)	6	72,473,700	73,272,164
(2) Filter 1	**(a)**	6	72,069,674	72,913,684
	(b)	6	45,153,840	45,981,562
(3) Filter 2	**(a)**	6	72,069,674	72,913,684
	(b)	6	45,153,840	45,981,562
(4) Dataset 1 (Chapter 4)	**(a)**	6	45,216,251	48,752,176
	(b)	6	70,581,495	73,639,640

Table 7. Positions of regions identified on BTA6 from the RH mapping using imputed data (http://www.ncbi.nlm.nih.gov/): (1) when using HD genotypes directly genotyped for Population 1 and imputed for Population 2; (2) after removing imputed genotypes for Population 2 with $r^2<0.7$; (3) after removing imputed genotypes for Population 2 with $r^2<0.9$. (4) Positions obtained without imputation and results from RH mapping as described in Chapter 4 (section 4.3.2) for Dataset 1. In the table, (a) and (b) represent the 1st and 2nd significant regions according to LRT.

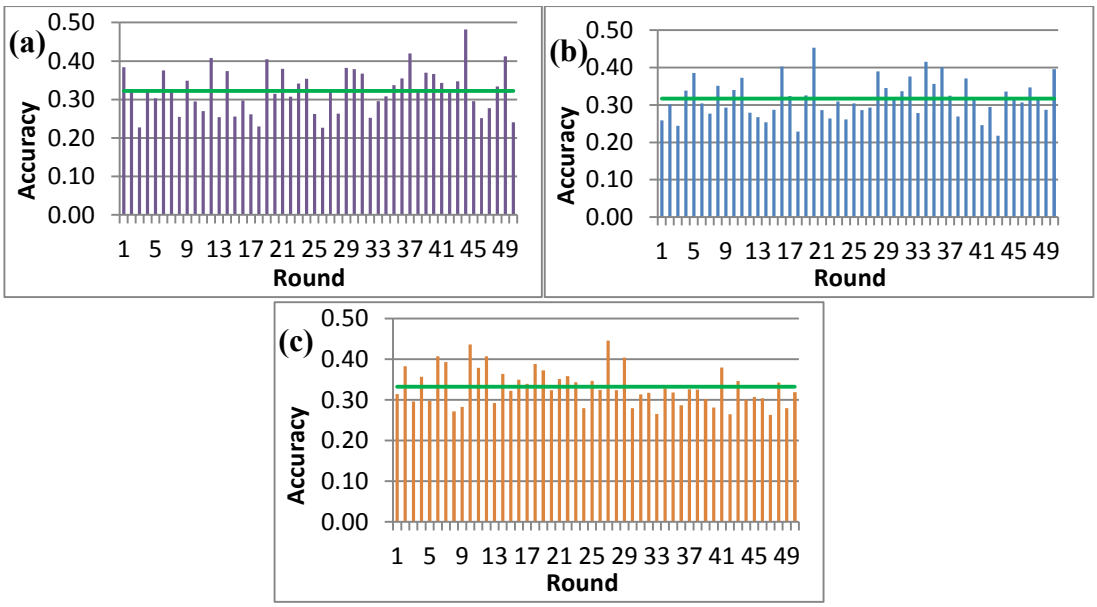


Figure 5. The accuracy of predicting the phenotypes after imputation for Population 2 from *50* Cross Validation randomisations when (a) using all the imputed genotypes, (b) after discarding all SNPs with $r^2<0.7$ (Filter 1) and (c) after discarding all SNPs with $r^2<0.9$ (Filter 2). The green lines represent the average across all *50* repeats.

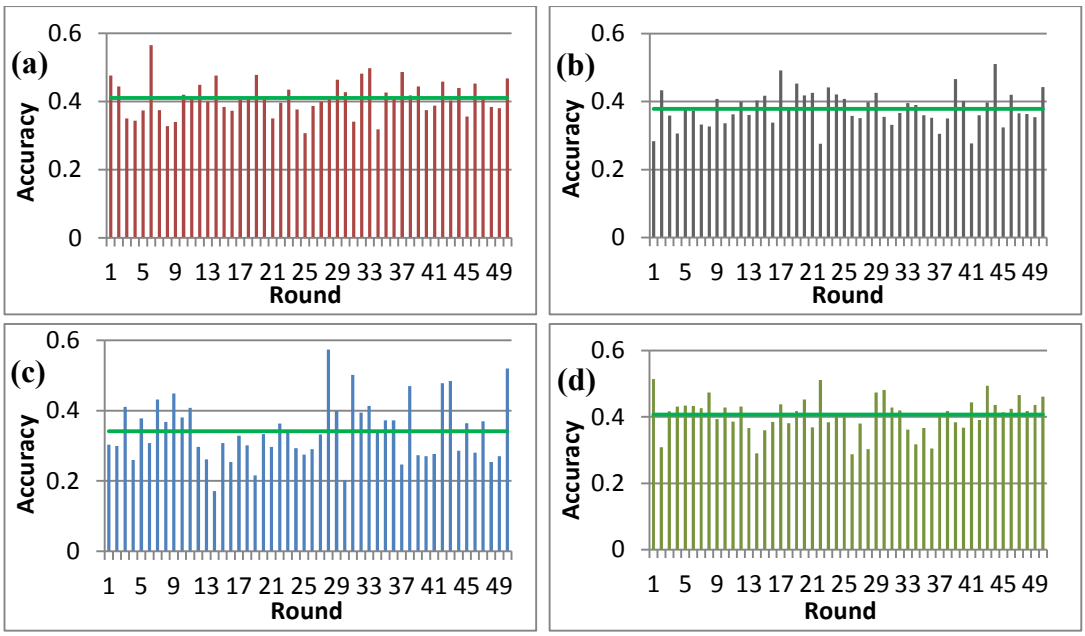


Figure 6. The accuracy of predicting the phenotypes after imputation for Population 2 from *50* Cross Validation randomisations when using (a) imputed genotypes only for BTA6 (after applying Filter 2) and the overlapping HD and low density SNPs for the rest of the genome, (b) imputed genotypes (after applying Filter 2) only for BTA8 and the overlapping HD and low density SNPs for the rest of the genome, (c) imputed genotypes (after applying Filter 2) only for BTA13 and the overlapping HD and low density SNPs for the rest of the genome, and (d) imputed genotypes were used simultaneously for BTA6 and BTA8 and the overlapping HD and low density SNPs for the rest of the genome.

	Average cor	Average accur	SE
Overlapping SNPs	0.116	0.33	0.050
All imp	0.108	0.323	0.058
Filter 1	0.104	0.317	0.052
Filter 2	0.110	0.332	0.046
Overlapping SNPs & Chr6 imp	0.131	**0.410**	0.052
Overlapping SNPs & Chr8 imp	0.105	0.379	0.052
Overlapping SNPs & Chr13 imp	0.082	0.341	0.086
Overlapping SNPs & Chr6 and 8 imp	0.129	0.407	0.054
Gen $_{-\text{Chr6}}$ + Chr6 imp	0.125	**0.368**	0.050

Table 6. Cross Validation prediction accuracies summary table from analysis on the combined directly genotyped and imputed genotypes, and after applying filters.

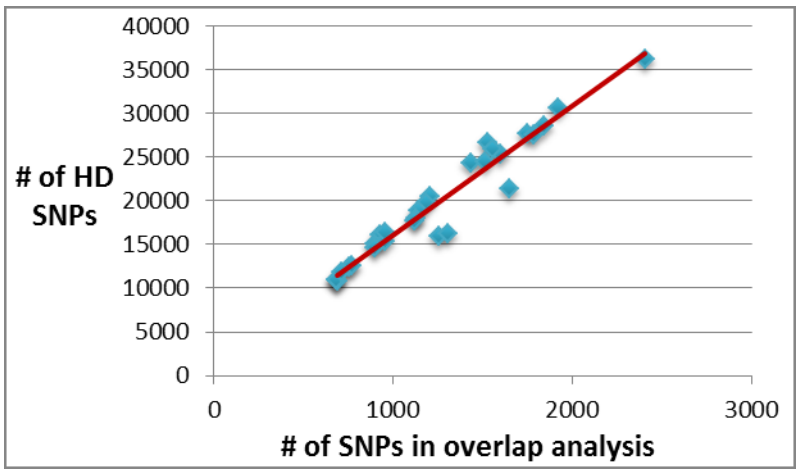


Figure 7. Regression of the number of HD SNPs per chromosome on the number of SNPs per chromosome appearing in both the HD and low density SNP Chips and retained in the analysis presented in Chapter 4 (sections 4.2.1.3 and 4.2.1.5).

Chapter 6

A comprehensive quantitative genetic analysis of the bTB diagnostic skin test SICCT

6.1 Introduction

BTB diagnosis in the UK has traditionally relied on the Single Intradermal Comparative Cervical Test (SICCT) (de la Rua-Domenech et al. 2006). The test, because of its comparative nature, allows differentiating true bTB infection from false positives due to exposure to *M. avium sbsp. avium,* and has a very good specificity (*>99% Sp*), however, it has a relatively poor sensitivity (*~55-70% Se*) (Neil et al. 1994; Olea-Popelka et al. 2004; De la Rua-Domenech et al. 2006).

Previous studies have demonstrated the feasibility of genomic selection for bTB resistance by (a) showing that there is heritable genetic variation for this trait (Bermingham et al. 2009; Brotherstone et al. 2010), and by (b) providing initial estimates of the accuracy for genomic selection (Tsairidou et al. 2014). Genomic selection of cattle for increased resistance to bTB might assist in the control of bTB. However, such selection will be partially informed by SICCT-based diagnosis of infection. Further, given the central role of SICCT in bTB control in the UK, it is important to know what would be the impact of genomic selection for bTB resistance on the SICCT test i.e. whether, in addition to increasing bTB resistance, this might also genetically alter actual SICCT values in both infected and uninfected cattle and, if it does, the likely magnitude of change. Therefore, understanding the genetic basis of variation in SICCT response conditional on the health status is important.

Genetic analyses previously conducted have demonstrated that resistance to bTB in first and in repeat breakdowns has a strong and positive genetic correlation, and although skin thickness itself was found to be highly heritable, there was strong indication that the heritability of the skin test is very low (DEFRA Evidence Project Final Reports SE3040 (2008) and SE3042 (2012)). Utilising field data in quantitative analyses is a challenging process, especially when identifying infected and healthy individuals relies upon imperfect diagnostics (Bishop and Woolliams 2010). During bTB testing, individuals that are non-reactors to the SICCT are classified as healthy, however, this class may also contain some misclassified bTB cases due to the imperfect sensitivity of the SICCT; reactors to the SICCT are classified as cases and they most likely have not been confirmed by other diagnostic means. Repeated measurements are available only for the animals classified as non-reactors or inconclusive reactors, while the reactors get a unique measurement and are immediately culled. Additionally, the presence of animals that are initially classified as healthy but become reactors at a repeated SICCT, raises questions on whether these animals may have been infected all the time and were false negatives due to the imperfect sensitivity of SICCT. Furthermore, SICCT records are collected by routine testing across the UK and the herds might either be undergoing a bTB breakdown or there might not be a confirmed breakdown, in which case the exposure status to bTB of the animals is unknown. The herds undergoing a breakdown might be in their first or later bTB breakdowns. All these difficulties in data structure introduce further challenges in the modelling of this data.

The aim of this study is to do a thorough quantitative genetic analysis of actual SICCT values, collected during bTB herd testing. Genetic variation analyses

have been based so far on confirmed cases and not on the SICCT itself. This chapter describes the modelling of SICCT utilising field data. The analysis that will be presented here aims to address an important preliminary question, namely the extent to which SICCT values are heritable in healthy or diseased cattle. Genetics can offer an effective tool for the control of bTB in cattle and since selection for bTB resistance will be based on diagnosis of the infection using the SICCT, this study investigates if such selection is likely to have an impact on the SICCT.

6.2 Materials and Methods

The genetic control of the response to the test was explored by means of fitting linear mixed models in ASReml. Data analysis comprised (a) heritability estimation analyses using SICCT data as collected, (b) heritability estimation after an appropriate transformation to normalise the residuals, (c) heritability estimation after attempting to remove the component of resistance to bTB (*R*) by taking into account the health status according to the indicator trait (i.e. reactor or not), (d) investigation of the impact of age on the heritability estimates, and (e) investigation of the correlation between the health status and the magnitude of response to SICCT.

6.2.1 Description of data

6.2.1.1 Data description and derived traits

The dataset comprises *117,356* Holstein-Friesian female cattle with *130,626* test records originating from *646* herds. Repeated measurements were available for *11,910* animals (*10,678* animals with *2* records, *1,104* with *3* records, and *128* with *4* records) that were non-reactors (NR) or inconclusive reactors (IR). All animals had

been tested over a period of *9* years (*2002-2010*), from herds undergoing their 1^{st}, 2^{nd}, 3^{rd} or 4^{th} bTB breakdown with the majority of animals being tested during the 1^{st} breakdown within a herd ($n= 112,116$) (Fig. 1). The age of the animals tested ranges from *43* to *6,605* days, with a mean of *1,397* days.

Four skin thickness measurements were available: skin thickness in millimetres at the site of avian tuberculin injection before inoculation (a_1), skin thickness in millimetres at the site of bovine tuberculin injection before inoculation (b_1), skin thickness in millimetres at the site of avian tuberculin injection after inoculation (a_2), and skin thickness in millimetres at the site of bovine tuberculin injection after inoculation (b_2) (Fig. 2). Three derived traits were of interest (i) $da=a_2-a_1$, which is the responsiveness to *M. avium*; (ii) $db=b_2-b_1$ which is the responsiveness to *M. bovis*; (iii) the skin test defined as *SICCT = db-da.* A detailed description of the data is presented in Table 1. The total number of non-reactors (NRs), inconclusive reactors (IRs), and reactors (Rs) according to (a) the standard interpretation (i.e. *SICCT<1mm* → NR, *SICCT=1-4mm* → IR, and *SICCT>4mm* → R), and (b) the severe interpretation (*SICCT<1mm* → NR, *SICCT=1-2mm* → IR, and *SICCT>2mm* → R) (Morrison et al. 2000) are presented in Figure 3. Due to the distribution of *SICCT* in this data with *~86%* of the *SICCT* values being zero, a derived *SICCT* (*logSICCT*) was calculated as $log_{10}(SICCT)+1$, after setting all $SICCT \leq 0$ values to *0.1*. All values were made positive by adding *1*, which did not make any difference to the ASReml analysis. The transformed data had a variance of *0.16* and a standard deviation of *0.4* (Fig. 4).

6.2.1.2 Data cleaning

This data had previously been cleaned to remove all records satisfying the following criteria: cows born before 1990; cows with a test carried out with a negative age or an age of ≤*42* days; male animals after sex validation; and cows with no sire recorded (Brotherstone S., personal communication, October 18, 2013).

Additionally in the present analysis, *2* animals with extreme a_2 and b_2 measurements of *77.00* and *99.00* respectively (considered to be arbitrary entries, either wrong or missing) were removed. According to the standard interpretation and using the individual a_1, a_2, b_1, and b_2 measurements in the data, *2* herds were found to have no standard Rs and less than *2* IRs, and *9* herds were found to have no Rs and no IRs (in total *418* records). For this analysis those herds were retained in the data. Further, based on the year of testing and the three tests available in the data for each record, *12* animals (*25* records) appeared to have been re-tested after being diagnosed as reactors. According to the re-testing protocol, once an animal is diagnosed as a reactor it is immediately culled, and therefore, these *12* animals were removed from the data. This left, *130,599* records for *117,342* animals that were retained in subsequent analyses.

6.2.1.3 Pedigree exploration

The animals in the data were offspring of *7,714* sires, out of which *5,510* sires had more than one daughter and the number of daughters per sire ranged with Q_0= Q_1=*1,* Q_2=*3,* Q_3=*9,* and Q_4=*2,028* where the *Q* values denote the quartiles (Fig. 5). Pedigree with both dam and sire known was available for *7,376* of those sires. Pedigree completeness was assessed by the number of equivalent generations

calculated using the ENDOG software (Gutiérrez et al. 2005), defined as the sum of $(1/2)^n$ terms over all known ancestors, where n is the number of generations separating the individual from the ancestor. The number of equivalent generations was estimated to be *4.27* in the pedigree provided.

6.2.2 Description of heritability estimation analyses

6.2.2.1 Preliminary analysis

(a) Analysis was conducted on all the *130,626* records. Heritability estimates were obtained in ASReml (Gilmour et al. 2002) for the three derived traits *SICCT*, *da* and *db*. The herd (*644* d.f.), test year (*8* d.f.), test month (*11* d.f.), and the interaction between test year and test month (*62* d.f.) were identified as fixed effects. The age and an indicator variable to account for the sequence of breakdowns within herd were included as covariates. The interaction of herd, date of breakdown and lactation group and the effect of sires were fitted as random effects. The pedigree file was used to account for relationships between sires. A sire model was fitted in ASReml as follows:

$$\boldsymbol{y} = \mathrm{m}\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathrm{W}\mathbf{h} + \mathrm{Z}\mathbf{u} + \mathbf{e} \quad (1)$$

where $\boldsymbol{y}$ is the response variable (i.e. *SICCT*, *da* or *db*), m is the overall mean, $\boldsymbol{1}$ is a vector of ones, $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{h}$ is the vector of the interaction term of herd, date of breakdown and lactation group fitted as a random effect with $\mathbf{h} \sim \mathrm{MVN}(0, \mathbf{I}\sigma_h^2)$, $\boldsymbol{u}$ is the vector of random sire effects with $\mathbf{u} \sim \mathrm{MVN}(0, \mathbf{A}\sigma_s^2)$, $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{Z}$ are the incidence matrices, and $\boldsymbol{e}$ is the residual error with

$\mathbf{e} \sim$ MVN $(0, \mathrm{I}\sigma_e^2)$. Heritability for the sire models was calculated as $4\sigma_s^2/\sigma_p^2$, where $\sigma_p^2 = \sigma_s^2 + \sigma_e^2$.

Locally Weighted Regression (LOESS) analysis was conducted on age and the *SICCT*, *da*, and *db* in order to assess the effect of age on the positive outcome of the skin test i.e. in the reactors (Fig. 6). In order to capture the age-related differences that were observed in the models, age was fitted as a polynomial. To select for the appropriate polynomial order for age, ANOVA analyses were conducted (a) for the *SICCT* and (b) for the *da*, where the herd, the test year, the test month, and the interaction between the test year and month were fitted as factors, and the age was fitted as a polynomial. Using the Akaike Information Criterion (AIC) the second order polynomial was selected for age in the *da* model. For consistency, the second order polynomial was also used for the *SICCT* model, despite the linear term appearing sufficient for *SICCT* based on AIC (Table 2).

Furthermore, variations in (a) were explored. Firstly, smoothing splines for age were fitted following model 2 below:

$$\boldsymbol{y} = \mathrm{m}\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathrm{f}(\mathrm{age}) + \mathrm{W}\mathbf{h} + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (2)$$

where $\boldsymbol{y}$ is the response variable (i.e. *SICCT*, *da* or *db*), m is the overall mean, $\boldsymbol{1}$ is a vector of ones, $\boldsymbol{\beta}$ is the vector of fixed effects, *f (age)* is a cubic spline with smoothing parameter calculated using ASReml and included in the random effects in ASReml, $\boldsymbol{h}$ is the vector of the interaction term of herd, date of breakdown and lactation group fitted as a random effect with $\mathbf{h} \sim$ MVN $(0, \mathbf{I}\sigma_h^2)$ and $\boldsymbol{u}$ is the vector of random sire effects with $\mathbf{u} \sim$ MVN $(0, \mathbf{A}\sigma_s^2)$, **X, W** and **Z** are the incidence matrices, and $\boldsymbol{e}$ is the residual error with $\mathbf{e} \sim$ MVN $(0, \mathrm{I}\sigma_e^2)$. Additionally, the year of

the test in days starting from a reference date (the minimum test year i.e. *2002*), and calculated as *tyrd=(tyr-2002)*365*, was fitted as a covariate replacing the test year. For a detailed presentation of the models used see Appendix 6.1a.

(b) Secondly, the analysis was repeated on the *130,599* records retained after data cleaning (section 6.2.1.2). Heritability estimates were obtained for the four derived traits *SICCT*, *logSICCT*, *da* and *db*. Further, in addition to the interaction term, the seasonality effect was also taken into account by calculating the "day within year" as *season=((tmn - 1)30 + tday)* and including that in the fixed effects as a covariate.

As described above, LOESS analysis for the reactors showed a distinct pattern for animals younger than *3* years, deriving from *db* (bovine tuberculin) (Fig. 5). Therefore, and in order to assess the effect of age, three contemporary age groups were distinguished as follows: Group 1: *age≤ 750d*, Group 2: *age > 750d* & *age ≤ 1100d*, and Group 3: *age > 1100d*) (Table 3). These age groups approximately correspond to different management groups for a *~25* months average age at first calving, and a voluntary waiting period of *~50* days, plus *~30* days until second successful conception (age at second calving *~1100* days). A new composite variable with the contemporary age group was fitted as a random effect (i.e. interaction of herd, date of breakdown and contemporary age group) which replaced the interaction of herd, date of breakdown and lactation group previously fitted as a random effect (see Appendix 6.1b).

6.2.2.2 Comprehensive analysis

This analysis was conducted on the *130,599* records retained after data cleaning. Heritability estimates were obtained in ASReml for the four derived traits *SICCT*, *logSICCT*, *da* and *db*. The herd and the interaction between test year and test month were fitted as fixed effects. The "season" term was not considered in the model as it is likely to be redundant since the interaction term between the year of the test and month of the test already captures seasonality. Age was fitted as a second order polynomial and the breakdown within herd was included as a covariate. The interaction of herd, date of breakdown and contemporary age group as described in 6.2.2.1 (b), and the effect of sires were fitted as random effects. Pedigree was used to account for relationships between sires and heritability for the sire models was calculated as previously. The general form of the models used was as in model (1) (for a detailed description of the models see Appendix 6.2). Further, analyses were repeated after additionally fitting a smoothing spline for age as in model (2) (see Appendix 6.2).

A hypothesis of interest is that the heritability that we detect is in part due to the health status classification of the individual (i.e. if it is a reactor or not) and it does not entirely correspond to genetic variation controlling response to the test. Therefore, in order to investigate the impact on the heritability after removing the resistance (*R*) component, the health status (*S*) was additionally fitted as a fixed effect having two levels: $S=1$ if $SICCT>4$ (i.e. for the Rs), and $S=0$ if $SICCT\leq 4$ (i.e. for the IRs and NRs).

6.2.2.3 Across-ages analyses

The aim of this analysis was to investigate the impacts of age on the heritability of the traits of interest and test whether the heritability estimates change with age. Following the three different age-groups observed (see 6.2.2.1 (b)), the data was subdivided by age in three subsets (Table 3), each subset containing only records of animals of age corresponding to that age-group. Heritability analyses were conducted in ASReml following the sire models described above, for *SICCT*, *logSICCT*, *da* and *db*, within each of the three age-groups. Furthermore, the analysis was repeated as a series of bivariate analyses between the different age-groups in order to obtain genetic correlations for the *SICCT* in the different age-groups.

Subsequently, the health status was additionally fitted as a fixed effect, as described in 6.2.2.2. As ASReml did not reach convergence, an alternative approach was followed to obtain correlations between the different age-groups using the EBVs for the sires obtained from the univariate analyses within each age-group and following the same model. This approach provides only approximate values for the genetic correlations as these EBVS are shrunk depending on the amount of information available for each sire, i.e. its reliability, and the less information is available the more the correlations obtained will be influenced by the environmental variance.

6.2.2.4 First records analyses

The aim of this approach was to obtain heritability estimates for a_1, b_1, and the derived traits of interest *SICCT*, *da* and *db*, after removing the repeated records

and retaining in the data only the first record for the animals that had repeated measurements.

(a) The first record was identified for each animal using the exact test date in days from a reference date (the earliest date observed in the data i.e. 14.10.2002), taking into account the leap years and the exact number of days in each month. The constructed dataset comprised *117,342* records. Moreover, the breakdowns were re-numbered continuously across herds (*728* new breakdowns) and the new breakdown describing every new mini epidemic was fitted as a fixed effect. The exact test date as described above and the age were fitted as cubic splines, and the sire was fitted as a random effect, as follows:

$$\boldsymbol{y} = \mathrm{m}\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathrm{f}_1(\mathrm{age}) + \mathrm{f}_2(\mathrm{date}) + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (3)$$

where ***y*** is the response variable (i.e. *SICCT*, *da* or *db*), *m* is the overall mean, ***1*** is a vector of ones, ***β*** is the vector of fixed effects, f_1 *(age)* is a cubic spline for age and f_2*(date)* is a cubic spline for the test date with smoothing parameters calculated using ASReml and included in the random effects in ASReml, ***u*** is the vector of random sire effects with $\mathbf{u} \sim \mathrm{MVN}(0, \mathbf{A}\sigma_s^2)$, **X** and **Z** are the incidence matrices, and ***e*** is the residual error with $\mathbf{e} \sim \mathrm{MVN}(0, \mathrm{I}\sigma_e^2)$. The splines used were either with *50* or with *100* knots. These analyses were repeated after additionally fitting the health status as a factor with two levels (see Appendix 6.3a).

(b) In the second part of this approach the dataset was further reduced to contain only the first known test within each of the new breakdowns and ignoring later tests. *88,932* records were retained in this analysis. This removed any information on the effect of the date of the test as the results from fitting models with

age, had indicated some cryptic structure in repeated tests giving unreasonable results for the effect of date on *SICCT* values (Fig. 8). The new breakdown was fitted as a fixed effect and a smoothing spline was fitted for age. The sire was fitted as a random effect. The sire models were fitted for a_1, b_1, *da*, *db*, and *SICCT* as follows (see Appendix 6.3b):

$$\boldsymbol{y} = \mathrm{m}\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathrm{f}(\mathrm{age}) + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (4)$$

where ***y*** is the response variable (i.e. *SICCT*, *da* or *db*), *m* is the overall mean, ***1*** is a vector of ones, ***β*** is the vector of fixed effects, *f (age)* is a cubic spline for age with smoothing parameters calculated using ASReml and included in the random effects in ASReml, ***u*** is the vectors of random sire effects with $\mathbf{u} \sim \mathrm{MVN}(0, \mathbf{A}\sigma_s^2)$, **X,** V, and **Z** are the incidence matrices, and ***e*** is the residual error with $\mathbf{e} \sim \mathrm{MVN}(0, \mathrm{I}\sigma_e^2)$. The analysis was repeated after removing the reactors after the standard interpretation (i.e. *SICCT>4*). This left *87,671* records and the same models as described above were fitted in ASReml. Lastly, these analyses were repeated after removing the reactors under the severe interpretation (i.e. *SICCT>2*). *86, 893* records were retained in the analysis.

Further, bivariate analyses were conducted between a_1 and b_1, a_1 and a_2, b_1 and b_2, *da* and *db*. Phenotypic, genetic and environmental correlations were obtained as well as regression slopes of b_1 on a_1, a_2 on a_1, b_2 on b_1, and *db* on *da*.

Additionally, the analysis on the first known test within each of the new breakdowns, was conducted within each of the three age-groups as defined above, following model (4). A multivariate analysis was conducted to obtain genetic correlations across the different age-groups.

Lastly, a bivariate analysis was conducted between the health status under the standard interpretation and *SICCT* to obtain a genetic correlation between the *SICCT* outcome and the healthy state (i.e. NRs and IRs). In this analysis the animals classified as reactors under the standard interpretation are treated as with missing *SICCT* measurement, while the non-reactors and the inconclusive reactors had their corresponding *SICCT* value. Further, this analysis was repeated after additionally removing records of *SICCT <-4*. This left *86,041* records in the analysis.

6.2.2.5 Supplementary analysis

The *SICCT* data calculated as described above, was analysed using an alternative approach. Following the model proposed by Hinger et al. (2008), two variables were created out of *SICCT*: *QPOS* and *QNEG*. In the *QPOS* variable the inconclusive reactors were included in the reactors (i.e. *QPOS*: *<1mm* → NR, and ≥ *1 mm* → R), while in the *QNEG* the IR were considered to be negative (i.e. *QNEG*: *≤4mm* → NR and > *4 mm* → R).

ASReml analyses were conducted either using a linear model for the *SICCT*, the *QPOS* ('TB status' as a binary trait) and *QNEG* ('TB status' as a binary trait), or using a threshold model and the logit link function for *QPOS* and *QNEG*. For all the approaches a sire model was used and tested either including relationship among sires or without considering relationship among sires. Age was fitted as a polynomial, the breakdown was fitted as fixed effect, the herd, test year, test month and test year by test month interaction were fitted as fixed effects, and the herd by breakdown date by lactation group interaction as well as the sire were fitted as random effects.

For the linear models, h^2 on the observed scale (h_o^2) was calculated using ASReml and then transformed to the liability scale using the formula $h_L^2=h_o^2\ p(1-p)/z^2$, with $z=pi$ (Robertson and Lerner 1949) where i is the mean deviation of individuals with values exceeding T and p is the prevalence defined as the proportion of infected animals in the sample.

6.3 Results

6.3.1 Preliminary analysis

Variance components and heritability estimates obtained from the different models used are presented in Table 4a. Heritabilities of *SICCT* were always less than *0.1*. Fitting a spline for age removed some of the variance explained by the sire component for the *SICCT* and for the *da* models, resulting in reducing the total genetic variance explained (Table 4a, *SICCT* (c) and (d), *da* (b)). The genetic variance for the *db* model was *0.186* (*SE=0.029*) and was larger than the genetic variance from all the *SICCT* models.

Variance components and heritability estimates obtained from the analysis following the additional cleaning of the data are presented in Table 4b. The heritability estimates obtained following data cleaning were slightly increased. The change in the heritability of *db* was greater compared to the other traits primarily due to an increase in σ_A^2. For *SICCT* and *da*, the changes in the estimates were within the range of their standard errors.

6.3.2 Comprehensive analysis

Variance components and heritability estimates obtained from the different models used are presented in Table 5a. These findings are in close agreement with the results from previous analysis reported in SE3042 (2012). The estimates corresponding to the heritability of the response to the test were very low, nevertheless detectable in data of this size.

When adding the observed health status as a fixed effect, which aims to remove the resistance component and thus may be closer to analysing the properties of the skin test, the heritability estimates for *SICCT* and *db* were reduced with both the genetic and the phenotypic variances being reduced. The heritability of *da* remained practically unchanged suggesting that the resistance component does not influence the measurements concerning reaction to the avian tuberculin.

Fitting a spline for age slightly reduced the heritability estimates obtained for all the traits (Table 5b). Similarly to what was observed above, accounting for the health status provided substantially lower heritability estimates for *SICCT* and *db* (Tables 5a and b, lower parts of the tables).

6.3.3 Across-ages analysis

ASReml heritability estimates for each of the three age-groups are presented in Tables 6a, 6b, and 6c. Heritability estimates for *SICCT* and *db* were increased for age-group *2* compared to age-group *1* or age-group *3* (Fig. 7). However, all the heritability estimates are very low and differences between the age-groups need to be assessed with respect to their standard errors. *SICCT* has a higher heritability in the

second age-group compared to the third age-group, while *db* seemed to follow a similar pattern being more heritable in the second age-group compared to the first and third age-groups. The *da* follows a different pattern, being more heritable in the first age-group, nevertheless providing more similar heritability estimates across all the age-groups. Heritability estimates for the *SICCT* and *logSICCT* followed similar patterns, with estimates for the log-transformed data being relatively inflated. Although a_1 and b_1 were highly heritable for young animals, their heritabilities reduced in the second and third age-groups. Similarly to what was observed in the comprehensive analysis of the full dataset, all heritability estimates obtained were reduced when the health status was fitted as a fixed effect.

Between the first and the second age-group, a genetic correlation of *0.43* (*SE =0.17*) was observed, while the third age-group had low genetic correlation with any of the other two groups (Tables 7a and 7b). Despite the volume of data, large standard errors are observed due to the low magnitude of the estimates of the heritabilities. When accounting for the health status, ASReml did not reach convergence and therefore, the correlations were obtained from the sire EBVs (Table 8). These correlations were larger than when not taking the health status into account (Tables 7a and 7b), however, the EBVs correlations only approximate the underlying genetic correlations.

6.3.4 First records analysis

(a) Analysis after retaining only the first record for each animal provided the estimates presented in Table 9, and the shape of the fitted values for the splines for age and date of the test can be seen in Figure 8. This analysis provided slightly

increased heritabilities compared to those obtained from analyses including the repeated records, however, the differences observed in the estimates were not significant given their standard errors, and all the estimates for *SICCT* were *<0.1*. Heritability estimates for a_1 and b_1 show that the skin thickness per se is a highly heritable trait (estimates in agreement with SE3042), with estimates for a_1 and b_1 being very similar. The heritability estimate for the *db* was higher compared to the estimate for the *da*. This *db* heritability includes the presence of genetic variance underlying bTB susceptibility and it would be expected to be captured by this model which does not take into account the health status and thus *db* heritability is estimated from data also containing truly infected individuals.

The models where the health status was additionally fitted as a fixed effect provided reduced heritability estimates for *SICCT* and *db*. However, taking the health status into account had no effect on a_1 or b_1, in agreement with the expectation as a_1 or b_1 are independent from infection. Further, fitting the health status had very little impact on *da*.

(b) Analysis after retaining only the first known test within each of the new breakdowns provided the estimates presented in Table 10. The results from the bivariate analyses between a_1 and b_1, a_1 and a_2, b_1 and b_2, *da* and *db*, can be seen in Table 13. The variances obtained in the bivariate analyses were similar to the estimates from the univariate analyses. The *SICCT* heritability was similar to previous estimates from analysis on the full dataset, and *db* provided an estimate similar to *da* as from analysis on full data. In the bivariate analyses a_1 and b_1 were highly genetically correlated with $cor_G(a_1, b_1) > 0.99$.

After removing the reactors under the standard interpretation, the estimates obtained from univariate and bivariate analyses are presented in Tables 11 and 14 respectively. Removing Rs had no impact on a_1 or b_1, and no impact on *da*. The heritability of the *SICCT* was slightly increased, while the heritability of *db* was slightly reduced compared to before removing reactors, but these differences were within the range of the standard errors of the estimates. Removing the reactors slightly increased the genetic correlations between a_1 and a_2, and b_1 and b_2. However, from the likelihood profile for the bivariate analysis of b_1 and b_2, it can be observed that the support interval for the genetic correlation is *0.93-0.98*, which is very high, but excludes a genetic correlation of *1*.

Results after removing the reactors under the severe interpretation, for univariate and bivariate analyses are presented in Tables 12 and 15. The heritability of *SICCT* was slightly increased and the heritability of *db* was slightly reduced, with the differences observed being within the range of their standard errors. The estimates for a_1, b_1, and *da* remained practically unchanged. The genetic correlation between a_1 and a_2 increased to *1*. The genetic correlation between b_1 and b_2 was *0.97*, which is increased, but is within the above support interval.

The results of the analysis within each of the three age-groups, when using the first known test within each of the new breakdowns, are presented in Tables 16a, b, and c. The second age-group provided higher heritability estimates for *SICCT* and *db* but not for the *da* (Fig. 9). Similar to what was observed in the across-ages analyses on all the records (i.e. including repeated records), although a_1 and b_1 are highly heritable in the first age group, the heritability becomes very small in age-groups *2* and *3*. The genetic variance for a_1 and b_1 changes dramatically with age,

however, by utilising *da* and *db*, this change is less important as the genetic variances for *da* and *db* are more consistent across age-groups.

The genetic correlations obtained from the multivariate analysis across the three age-groups can be found in Table 17. Due to the very low genetic variances, this analysis had low power for estimating between-age-groups correlations and the ASReml analysis provided a genetic correlation between age-groups *1* and *2* that was on the boundary. The log likelihood of this model was *-1223.17*. Under the null hypothesis that *SICCT* among the different age groups are not genetically correlated, the genetic correlations for all three traits was fixed to *0* and the new likelihood was *-1225.16* $(LRT=2*(L_1-L_0)=2*(-1223.17 - (-1225.16)) = 3.98$, which is not significantly different than the H_0 ($X^2=7.815$, *for 3 d.f.* $p=0.05$). Thus, zero cannot be excluded and the correlation between the different age-groups is possible to be zero. Conversely, the genetic correlation between all the age-groups was fixed to be 0.99, and this model provided a likelihood of *-1226.35* and *LRT=6.36*, again not significantly different. These results indicate that for this analysis, there is very little information in the data.

The bivariate analysis between the healthy status and *SICCT* provided a genetic correlation of *-0.01(SE = 0.14)*. This correlation can be interpreted, with caution, as the genetic correlation of the *SICCT* in healthy individuals, with susceptibility to disease. The caution arises from the identification of "healthy" with not being a standard reactor and reactor status as a measure of disease. This value close to zero indicates that the magnitude of response to the *SICCT* in healthy individuals is not genetically correlated with susceptibility to disease. The bivariate analysis between the health status and *SICCT* when additionally removing the

records with *SICCT <-4* provided again a very small genetic correlation with a value of *0.07 (SE = 0.16)*.

6.3.5 Supplementary analysis

The results of this analysis can be found in Tables 18 and 19. Estimates of the heritability on the observed scale (h^2_O) and on the liability scale (h^2_L) were provided from the ASReml analysis for the linear and the threshold models respectively (Table 18). For the estimates to be comparable h^2_O was transformed to the liability scale, however, h^2_L was highly dependent on the definition of prevalence used, i.e if it had assigned values assumed for the population ($p=10\%$ *or* $p=6\%$) (Table 18), if it was the proportion of reactors in the entire sample (p_{SICCT}), or if the apparent p was used (i.e. *QPOS* and *QNEG* specific p, as whether the IRs are considered to be bTB negative or positive changes the prevalence, with $p_{QNEG}=p_{SICCT}$) (Table 19).

QPOS and *QNEG* categorisations provided very different h^2_L estimates from the linear models depending on the assumed population prevalence. This difference was smaller when the apparent p was used. Comparing threshold models and linear models for the binary traits (i.e. *QPOS* and *QNEG*), the threshold model for *QPOS* provided lower estimates than the linear model, while for *QNEG* the threshold model provided greater estimates than the linear model. For the binary *QPOS* and *QNEG*, the sire models including relationships between sires provided greater estimates compared to sire models without relationships between sires, as it has been previously observed in the analysis by Hinger et al. (2008) on Paratuberculosis data. The estimates (h^2_L) provided from the linear models for the *SICCT* when it was

considered as a linear trait, were generally lower than the estimates when it was considered as a binary trait (*QPOS* or *QNEG*).

6.4 Discussion

6.4.1 The genetics of SICCT

For a bTB control strategy based on genetics and informed at least in part from the *SICCT*, one of the critical questions arising is what the likely impacts of such selection would be on the *SICCT* itself. In other words, when selecting for genetically bTB resistant animals based on the *SICCT*, a danger might be the unintentional selection for animals that respond less to the *SICCT*. This ambiguity haunts decision makers when it comes to bTB control. Therefore, understanding the genetic control of the response to the *SICCT* is crucial in designing a bTB control strategy. The extent to which genetics control the *SICCT* outcomes can be investigated through its heritability. In the present study, the genetics of *SICCT* were explored through the systematic modelling of *SICCT* data and through conducting a thorough variance component analysis on the *SICCT* itself. For this purpose, extensive field data was utilised collected during bTB breakdowns across UK herds.

6.4.1.1 Development of models

In this Chapter, alternative models were tested sequentially for the *SICCT* and its components *da*, *db*, a_1, b_1, a_2, and b_2. Initially, the interaction of herd, date of breakdown and contemporary age group was fitted as a random effect (Comprehensive analysis and Preliminary analysis). However, the herds might be either randomly distributed with no specific structure, or they might differ in a

systematic way. These systematic differences can be for example the geographic location of the herd or management differences e.g. the grazing system or the milk yield, which might affect the bTB prevalence in specific herds introducing some structure. Given the information available, the exact sources of such structure are unknown and thus cannot be directly accounted for. If the herd was fitted as a random effect assuming randomly distributed herds, the possibility of such structure would be ignored and re-interpreted by terms in the model including genetic terms, neglecting the possibility that hidden structure may be confounded to some degree with genetics. Therefore, because of the operational importance of the issue and to avoid any ambiguity, it was preferred to fit the herd as a fixed effect which allows recovering information within herds and avoids making assumptions on the nature of the herd differences.

In later models (First-records analysis), the herd was implied within the "new breakdown" variable which considers every breakdown in every herd as a new mini-epidemic. The "new breakdown" was fitted as a fixed effect to account for systematic differences among breakdowns and recover information within breakdowns but not between breakdowns.

The *logSICCT* provided slightly higher heritability estimates compared to the non-transformed *SICCT*, as the transformation influences the variance. By setting all $SICCT \leq 0$ values to 0.1, more weight is applied on the contrast between being a non-reactor (NR) and being a reactor (IRs and Rs), and thus more emphasis is applied on larger values, i.e. on the reactors and likely reactors.

Modelling bTB field data is further challenged by ambiguities in the records, such as measurement recording errors. The need for improving the quality of readings and the importance of better consistency across testings has been demonstrated in previous studies (Clegg et al. 2015), and has been recognised by the UK government by introducing quality control in bTB test recording. However, given the data currently available, the confidence in some of the conclusions drawn reflects the uncertain quality of the data. Reviewing and restructuring the testing and recording practices is one of the critical challenges that need to be met in bTB control.

6.4.1.2 The hierarchy of the test

The heritability of the different components of the test is a source of different types of information about the underlying genetics. In the UK, the comparative version of the tuberculin test, i.e. the *SICCT*, has been employed for the surveillance of bTB, in the belief that exposure to *M. avium sbsp avium* can generate false positive reactions. Following the studies by Lesslie et al. (1975), the comparative test has better specificity compared to its non-comparative equivalent, and thus, improved power to identify the true bTB case. However this assertion has not been adequately tested. One of the novelties in the present study is that the individual components of *SICCT*, are investigated at the genetic level aiming to provide a better understanding of the genetic background of the *SICCT*. More specifically, the hierarchy in the design of *SICCT*, i.e. going from four individual a_1, b_1, a_2, and b_2 skin thickness measurements, to $da=a_2-a_1$ and $db=b_2-b_1$ and ultimately *SICCT* derived as $db - da$, minimises the possibility that animals which are not bTB infected

but are exposed to *M. avium sbsp. avium* will be identified as reactors to the test (i.e. false positives) and, as it was demonstrated in the present study, this hierarchy is also of paramount importance at the genetic level. The interpretation of the genetic control of the individual components of the *SICCT* will be presented in the following paragraphs.

The a_1 and b_1 provide information on the skin thickness as a trait, and in the present study a_1 and b_1 were found to be highly heritable, as will be discussed below in more detail. The a_2 and b_2 are the skin thickness measurements after inoculation of the avian and bovine antigens respectively. Under the polygenic trait assumption, the genetic correlation informs on the strength of the relationship between the traits, and thus, what fraction of loci they have in common, while the regressions informs on how the scale of the response in the one trait relates to the scale of response in the second trait. In the first-records analysis and when the reactors are included, the a_1 and a_2 were positively correlated, and had a stronger correlation compared to b_1 and b_2 (Table 13). Removing the reactors increases the genetic correlation between b_1 and b_2, making the pattern observed between a_1 and a_2, and b_1 and b_2, more similar. From the regression of a_2 on a_1 it can be explored whether a_2 depends on a_1, and similarly for b_2 and b_1. These regressions were not much greater than *1*, which demonstrates that *da* and *db* are independent to a_1 and b_1. The "post-bovine antigen inoculation" and "post-avium antigen inoculation" are not being scaled by a_1 and b_1 i.e. *SICCT* does not depend on the initial skin thickness. Therefore, whether different genetic groups have different genetic variances for a_1, becomes less important as a_1 and a_2 are highly correlated with a regression of *1*, and consequently there is no

dependence on skin thickness. This is stressing the importance of the hierarchy within the *SICCT* and the value of using a_2-a_1 and not a_2 alone.

Further insight into the importance of the hierarchy of the test at the genetic level, can be gained if we consider the genetic variances captured by each of the *SICCT* components and what these variances represent. If using only a_2 and b_2, their genetic variance would capture the genetic variance of the skin thickness which is a highly heritable trait. Therefore including a_1 and b_1 and calculating *da* and *db* makes the test more stable with respect to the genetics of the skin thickness. The *da* and *db* provide information on the genetics of the responsiveness to the avian and the bovine tuberculins, separately from one another. While response to the *SICCT* refers to the magnitude of the *SICCT* outcome, responsiveness to the avian and bovine tuberculins captured by the *da* and *db*, refers to the ability of the immune system to respond to the antigens. For example in humans, reactivity to the non-comparative tuberculin skin test and immune response to *Mycobacterium tuberculosis* antigens with respect to IFN-γ, TNAa and antigen-specific cells production, has been shown to be moderately to highly heritable (Jepson et al. 2001; Stein et al. 2003; Stein et al. 2005; Cobat et al. 2010), while a linkage analysis has identified major loci associated with TB skin test reactivity (Cobat et al. 2009), and genetics have been linked to the outcome of infection (Stein et al. 2008). Therefore for bTB, using *SICCT* defined as *db-da*, is crucial not only for reducing the false positive reactions, but also at the genetic level. To demonstrate this further, although *da* and *db* are highly genetically correlated (Tables 13, 14, and 15), they both provided higher heritabilities than the *SICCT*. By employing *db-da*, the common genetic part of *da* and *db* is removed and thus, the heritability of *SICCT* is lower than its components. Therefore, the

comparative *SICCT* is a more stable test compared to a non-comparative equivalent at the genetic level. An important outcome is that if selection was based on *db* alone, there would be a greater risk of response to selection compared to the use of *SICCT* because of the higher heritability of *db*. Therefore, the analysis presented in this Chapter, confirmed at the genetic level the wisdom of the hierarchy of *SICCT*.

6.4.1.3 The impacts of the health status on *SICCT*

In this Chapter, an approach is presented that allows disentangling the genetics of bTB resistance from the genetics of response to the *SICCT*. The skin test is simply a measurement (i.e. an indicator trait), and not the trait, thus, as the heritability of resistance to bTB (*R*) is non-zero (h^2_O*=0.23 (SE=0.06)*, see Chapter 2), even if the heritability of the skin test itself in infected and uninfected animals was truly zero, we would still observe a non-zero heritability for the *SICCT* as it captures part of *R*.

This was addressed by fitting the health status as a fixed effect, therefore removing the *R* component from the genetics, which is closer to analysing the properties of the skin test. The *R* component is fitted as a factor with two levels corresponding to a healthy or a diseased health status without implying a genetic correlation between resistance and the magnitude of the *SICCT*. Conclusions are drawn within the classes of *R,* however, the genetic parameters of responses to *SICCT* in non-reactors are assumed to be the same as the genetic parameters of responses to *SICCT* in reactors. Additionally, how reliable is the health status needs to be considered with respect to the sensitivity of the *SICCT*. Since the classification in healthy and diseased is not confirmed, the health status *R* is solely based on

SICCT, and consequently some of the animals classified as healthy will be false negatives. Due to these weaknesses, this approach is a first step in taking into account the health status, but may be statistically and biologically naïve, and thus needs to be considered in more depth.

6.4.1.4 The impacts of age at test on the outcome of *SICCT*

The consistency of heritability estimates across the identified age-groups was examined to test whether the heritability changes with age. The *SICCT* and the *db* seem to follow a similar pattern being more heritable in the second age-group. This could indicate that in older cattle *SICCT* and *db* responses are more sensitive to management changes and less controlled by genetics. From the genetic correlation of *SICCT* across the different age-groups the likely selection intensity across different ages was examined. If *SICCT* is not genetically correlated between the different age-groups, then selection intensity arising from culling false positives, is different among different age-groups and any unintentional selection is diluted and possibly not in the same direction (Table 17). However, this data was not found to have enough information for obtaining reliable estimates of genetic correlations across the different age-groups.

In the previous report (DEFRA Evidence Project Final Reports SE3042, 2012), skin thickness itself was found to be highly heritable. Skin thickness is a characteristic that varies between different cattle breeds (Dowling 1955; Dowling 1963) and has been studied in relation to heat tolerance (Alfonzo et al. 2015) and tick resistance (Riek et al. 1962; Marufu et al. 2013), although neither study was conclusive about its importance to the respective trait. In the present study and when

retaining only the first record for each animal, a_1 and b_1 were shown to be highly heritable in the first age-group, however, the heritability of a_1 and b_1 was low in age-groups *2* and *3*. This possibly reflects that the skin thickness increases until maturity is reached (Dowling 1964). Growth in cattle is a heritable trait, with heritabilities varying between different breeds, and with heritability decreasing with increasing age (Lin et al. 1985; Groen and Vos 1995; Coffey et al. 2006). The animals in age-group *1* are still in development, thus, the rate of increasing their skin thickness is controlled by genetics i.e. how early-maturing they are is under genetic control. When they reach maturity their skin thickness reflects the breed standard and thus their genetic differences are reduced since all the animals in the data are of the same breed. Additionally, from the age distribution in the data (Fig. 6a) it becomes apparent that the majority of records in this data are from younger animals where the phenotypic variance is greater. Furthermore, when breaking down the data into age-groups, the heritability estimates for a_1 and b_1 are reduced in age-groups *2* and *3*, while the heritabilities for a_1 and b_1 are high in the entire data where the estimates are dominated by the most variable group, i.e. age-group *1*, which has the highest heritability for a_1 and b_1. Thus, although a_1 and b_1 estimates change quite dramatically over age-groups, using *da* and *db* minimises the age-group differences as the genetic variances of *da* and *db* are less influenced by age and are more consistent across age-groups, controlling better for some of this change.

Lastly, an important environmental confounder that currently is not taken into account by *SICCT*, is the Paratuberculosis status. Paratuberculosis caused by *M. avium sbsp paratuberculosis* emerges over a long period of time and interactions

between Paratuberculosis infection and Paratuberculosis vaccination status, with the *SICCT* might also be time-dependent.

6.4.2 Comparison with previous report

In the previous report (DEFRA Evidence Project Final Reports SE3042, 2012) there was strong indication that the *SICCT* has a very low heritability. In agreement with those findings, after fitting systematically a series of alternative models, the heritability of the skin test *SICCT* was detectable in data of this size but was very low for all the models tested.

More specifically, when using all the records (see Preliminary analysis (Tables 4a and 4b) and Comprehensive analysis (Tables 5a and 5b)), the heritability estimates obtained for the *SICCT*, the *da* and the *db*, were very close to the estimates presented in SE3042 (2012, Table 2). In addition to what was done in the previous report, this analysis considered the likely impact of the presence of repeated records for the animals that were not initially classified as reactors, and the presence of records from multiple breakdowns, on the heritability estimates. The issues arising from the repeated measurements in the data can be dealt with either fitting the individual animals as a random effect or removing repeated measurements and retaining only the first record for each animal in the data. However, in the present data some measurements are repeated after a few weeks and sometimes repeated tests might occur after years. Models that attempted to include repeated records provided unrealistic time-trends for the mean skin test response (Fig. 8), i.e. a *4mm* change in mean *SICCT* during the course of collection of this data, indicating a

weakness in the models adopted for the repeated records. Therefore, it was preferred to remove the repeated records from the data.

In the first-record analyses, the estimates obtained for the *SICCT* and especially for the *db*, were higher compared to the estimates from the analyses on the entire dataset (Table 9). For the *SICCT* and the *db*, having the repeated records in the data did not influence the error variance but it provided reduced estimates of the genetic variance. A possible explanation for this is that due to the structure of the data with only the healthy animals having repeated measurements, the healthy individuals get additional weight in the analysis on the full dataset as they are more often represented in the data. This results in truncating the distribution and diluting the estimates of the genetic variance, as this variance comes from the individuals without repeated records (i.e. the identified bTB infected animals). However, after keeping only the first test within each breakdown (Table 10), *SICCT* and *db* heritabilities were reduced with both the error and the genetic variances being reduced (Table 10).

In the first-records analyses removing the reactors slightly increased the heritability of *SICCT*, which can be attributed to the reactors being highly variable in this data, and therefore by removing them, a source of considerable variance is removed. The heritability increased due to changes in the phenotypic variance i.e. the residual variance. For *db*, the phenotypic variance decreased and consequently also the regression of *db* on *da* was reduced in this analysis.

6.4.3 SICCT and selection for bTB resistance

In the present study, response to the *SICCT* diagnostic test was found to be lowly heritable, indicating that any risk of unintentional selection for low response to the skin test is very small.

In a previous study (Amos et al. 2013), a microsatellite marker genotype was linked with reduced immunological reaction to tuberculins and was found to be more frequent among non-reactor animals, from which the authors concluded that there is a genetic predisposition for some animals to pass the standard *SICCT* although infected with bTB, and these animals are selected through the test-and-slaughter policy. However, that study was conducted on a mixed population of designated breeds and crosses with very small sample sizes for some breeds which would substantially reduce the power to detect associations, and the analysis was conducted without correcting for population structure which might lead to spurious associations. Microsatellite markers are less suitable for association analyses as they are highly polymorphic, and unless very dense, they are not likely to have high LD between the marker and the causative mutation, which is the basis of any association analysis. Especially in a multi breed analysis the same linkage phase across all the breeds would be required but it would be highly unlikely to occur unless the causative mutation was very close to the marker. Moreover in Amos et al. (2013), the phenotypes comprised records of animals that tested negative to the *SICCT* and thus there was no information on animals that were identified as reactors, which would limit the conclusions only within the class of non-reactors.

The importance of the hierarchy and the comparative structure of the *SICCT* has been highlighted above (section 6.4.1.1), and in the present study the genetics of *SICCT* as well as the genetics of its components *da*, *db*, a_1, a_2, b_1, and b_2 were explored. Although, Amos et al. (2013) modelled the components of *SICCT*, concluding that a certain genotype was associated with *da*, the *SICCT* itself was not included in the models. As it has been demonstrated in this Chapter, the genetic properties of *da*, *db* and *SICCT* differ. To extrapolate from *da* and *db* to *SICCT*, evidence on the genetic correlation of *da* with *db* is needed. It is only the *SICCT* that captures the comparative test result used for identifying reactors, and the heritability of *SICCT* was found to be very low, and lower than the heritability of *db*. Thus, the likely selection intensity to *SICCT* is very small, and smaller than the likely selection if we were using *db* alone.

The specificity of the test is very high, thus the test is removing only a very small proportion of healthy animals. Even in the unlikely case that there would be some selection pressure on *SICCT*, given its very low heritability, any response to such selection would be weak and highly unlikely to occur within a reasonable time period. Further, even if the test was heritable, selection would remove the extremes, and thus heritability of *SICCT* and subsequently response to selection on *SICCT* would further decline.

In Tsairidou et al. (2014), it was demonstrated the feasibility of genomic selection for bTB resistance based on confirmed cases, i.e. diagnosed with bTB through the *SICCT* and with lesions in the abattoir. The greatest impact of genomic selection informed by the *SICCT* will be to reduce the susceptibility of the general population to bTB infection. Given the very low heritability of *SICCT*,

implementing genetic selection for bTB resistance is unlikely to compromise the integrity of the test in the near future. Conceivably over time, some response to selection might occur in the response to *SICCT* and some individuals might be less responsive, i.e. might reduce the sensitivity of *SICCT* as infected animals might be less likely to be identified. Individuals carrying genotypes that might give responses to *SICCT* when diseased that are below the threshold for declaring a reactor, already exist in the population and already go undetected. Moreover, these genotypes are currently open to genetic drift, and chance selection of Holstein bulls carrying these genotypes might increase their frequency through drift. Given that at present there is no monitoring, there is no ability to assess these effects. Although the risks for unintentional selection on the response to *SICCT* are small, monitoring *SICCT* while conducting selection for bTB resistance, will allow any changes in the magnitude of response to be monitored, and if judged beneficial, the classification thresholds of *SICCT* can be adopted appropriately to provide the desired test sensitivity.

Given the very low heritability of *SICCT*, any such effects might only be observed after a long period of time, when there will already be effects of selecting for bTB resistance i.e. the average bTB resistance of the population will have been already increased and the bTB prevalence will be much lower. While increasing resistance and as the epidemic declines, it will be more difficult to detect bTB using *SICCT*, and bTB infected animals might become more difficult to identify. However, by increasing the average population resistance, R_0 will be eventually brought below *1*, so the risk of an epidemic will become negligible (Bishop 2010). Genetically susceptible cattle that might be more difficult to detect using *SICCT*, and thus remain in the population, will not be able to cause any major epidemics while any minor

epidemics that might start, will die out on their own. The issue of the environmental reservoir is discussed in the general discussion.

In the literature, it has been argued that selection using the skin test as a threshold for culling standard reactors, may change the properties of *SICCT* in healthy animals (Amos et al. 2013). In the analysis presented in this Chapter, in individuals classified as healthy, the magnitude of response to *SICCT* was not found to be genetically correlated with those that pass the threshold. This result indicates that selection for individuals less susceptible to bTB is not likely to change the magnitude of the response to *SICCT* in the healthy individuals. However, due to the imperfect sensitivity of *SICCT*, the group of healthy animals also contains the animals classified as healthy although they are diseased, i.e. the false negatives. Dealing with the imperfect sensitivity of the *SICCT* is another critical challenge for bTB control.

6.4.4 Conclusion

The cow's response to the *SICCT* skin test was found to be very lowly heritable, thus, genetic selection for bTB resistance is unlikely to compromise the integrity of the diagnostic test in any reasonable time period. Removing the component of resistance to bTB had no impact on a_1, b_1, or *da*, but reduced the heritability estimates for the *SICCT* and the *db*, showing that the estimated heritability for *SICCT* partially captures the component of genetic resistance to bTB and thus the true heritability of *SICCT* itself is even lower. Therefore, any risk of unintentional selection for low response to the skin test is very small, and it is the imperfect sensitivity of the test that it is more likely to be one of the critical factors

contributing to the difficulties in the control of bTB. Investigation of the impact of age on the heritability estimates showed that *SICCT* and *db* are more heritable in the second age group i.e. cows in their first lactation, while given the information available in this data, limited conclusions could be drawn about the genetic correlation of *SICCT* across the different age groups. Lastly, the magnitude of response to the *SICCT* was not found to be genetically correlated with the healthy status indicating that selection for individuals less susceptible to bTB is not likely to change the magnitude of response to the *SICCT* in healthy individuals. The hierarchy of *SICCT* was shown to be important both at the genetic and phenotypic level.

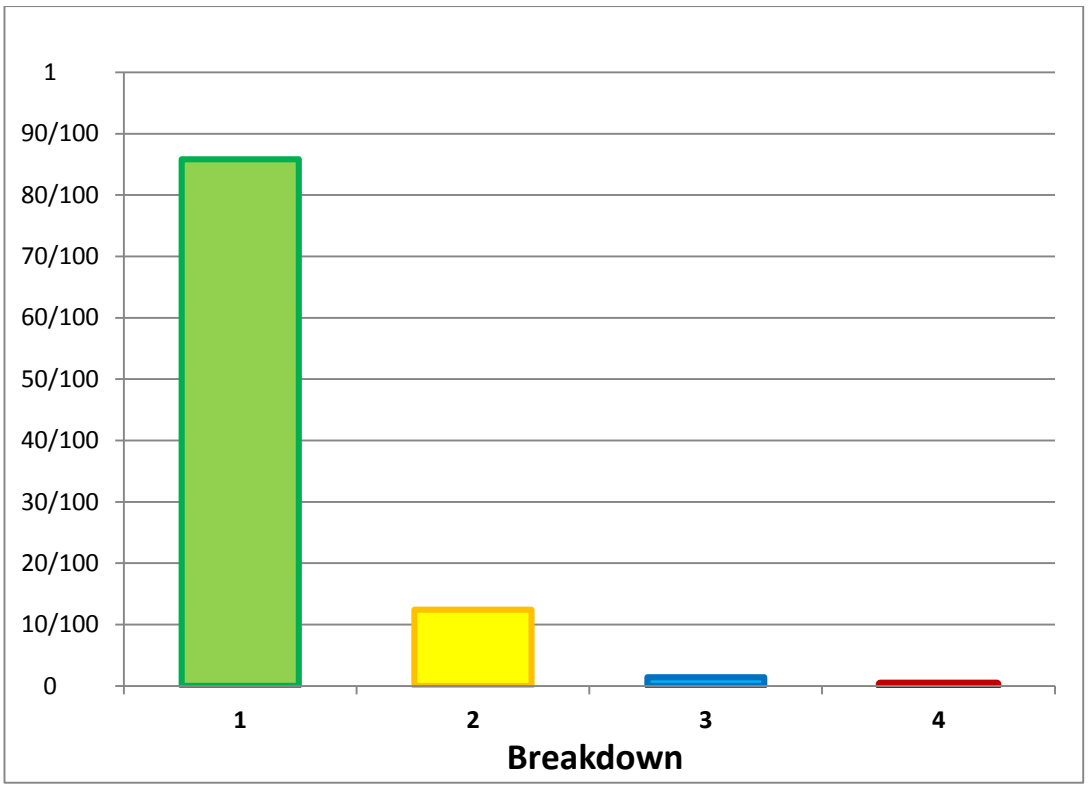


Figure 1. Percentages of the test records during the *1st*, *2nd*, *3rd* and *4th* breakdown within a herd.

	Min	Median	Mean	Max	Var	SD	Mode
a_1	1.00	6.00	6.17	16.00	2.33	1.53	6.00
a_2	0.00	6.00	6.83	50.00	6.31	2.51	6.00
b_1	1.00	6.00	6.19	16.00	2.33	1.53	6.00
b_2	0.00	6.00	6.97	90.00	10.46	3.23	6.00
da	-7.00	0.00	0.66	43.00	3.54	1.88	0.00
db	-7.00	0.00	0.78	81.00	7.53	2.74	0.00
SICCT	-43.00	0.00	0.12	81.00	5.72	2.39	0.00

Table 1. Descriptive statistics for the four skin thickness measurements, before (a_1 and b_1), and after (a_2 and b_2) inoculation of the tuberculin antigens, and for the derived traits $da=a_2-a_1$, $db=b_2-b_1$, and the skin test $SICCT=db-da$ following data cleaning.

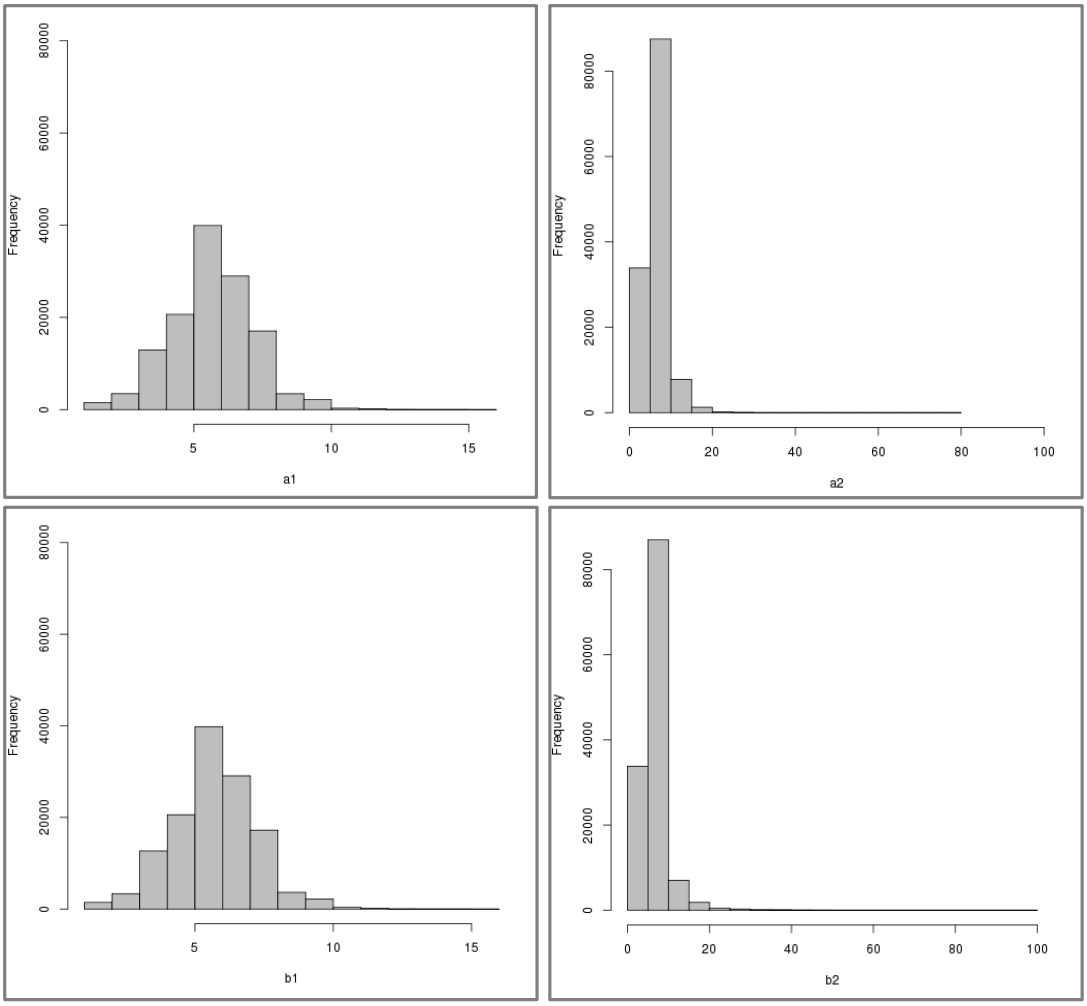


Figure 2. Histograms of a_1, a_2, b_1, and b_2.

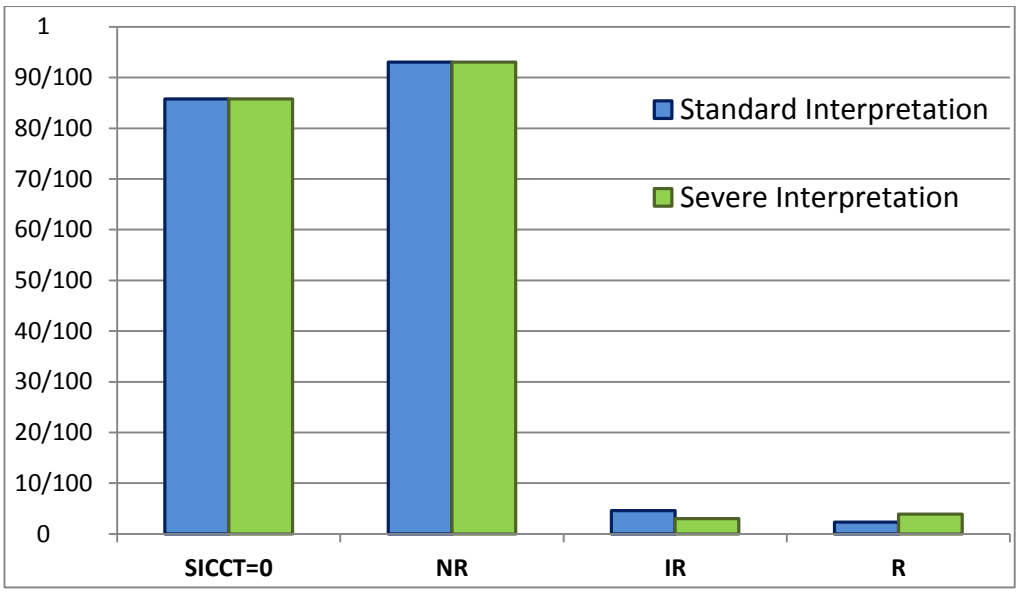


Figure 3. Fractions of non-reactors (NR), inconclusive reactors (IR), and reactor animals (R), according to the *SICCT* result after the standard and the severe interpretations.

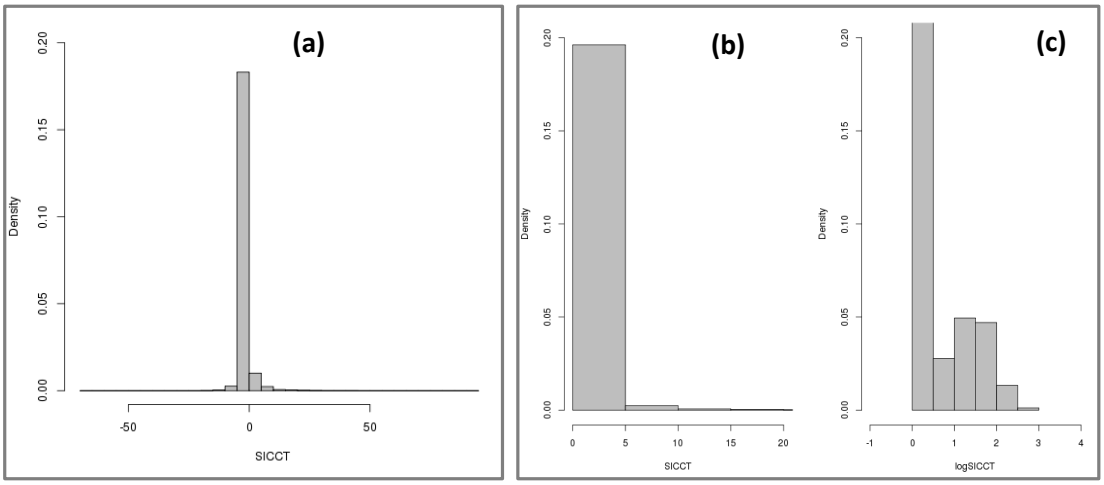


Figure 4. Distribution of *SICCT* results (a) before transformation, (b) after setting all values ≤*0* to *0.1* and (c) after using $log_{10}(SICCT)+1$ transformation.

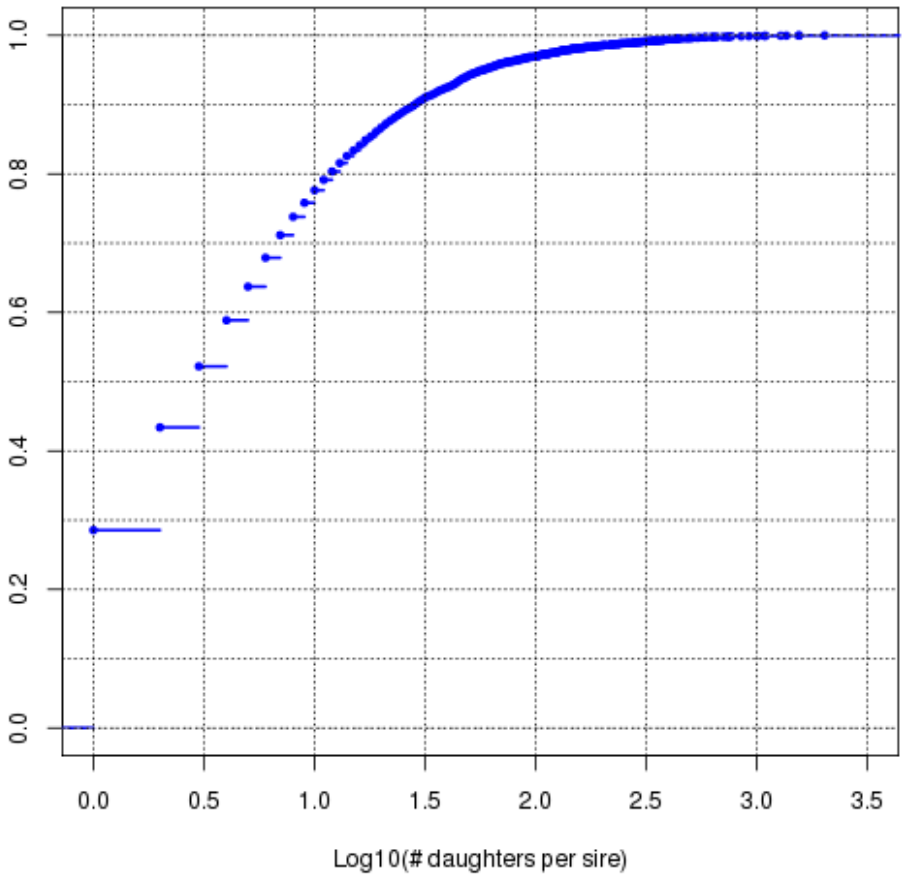


Figure 5. Cumulative distribution of the number of daughters per sire for the total number of sires.

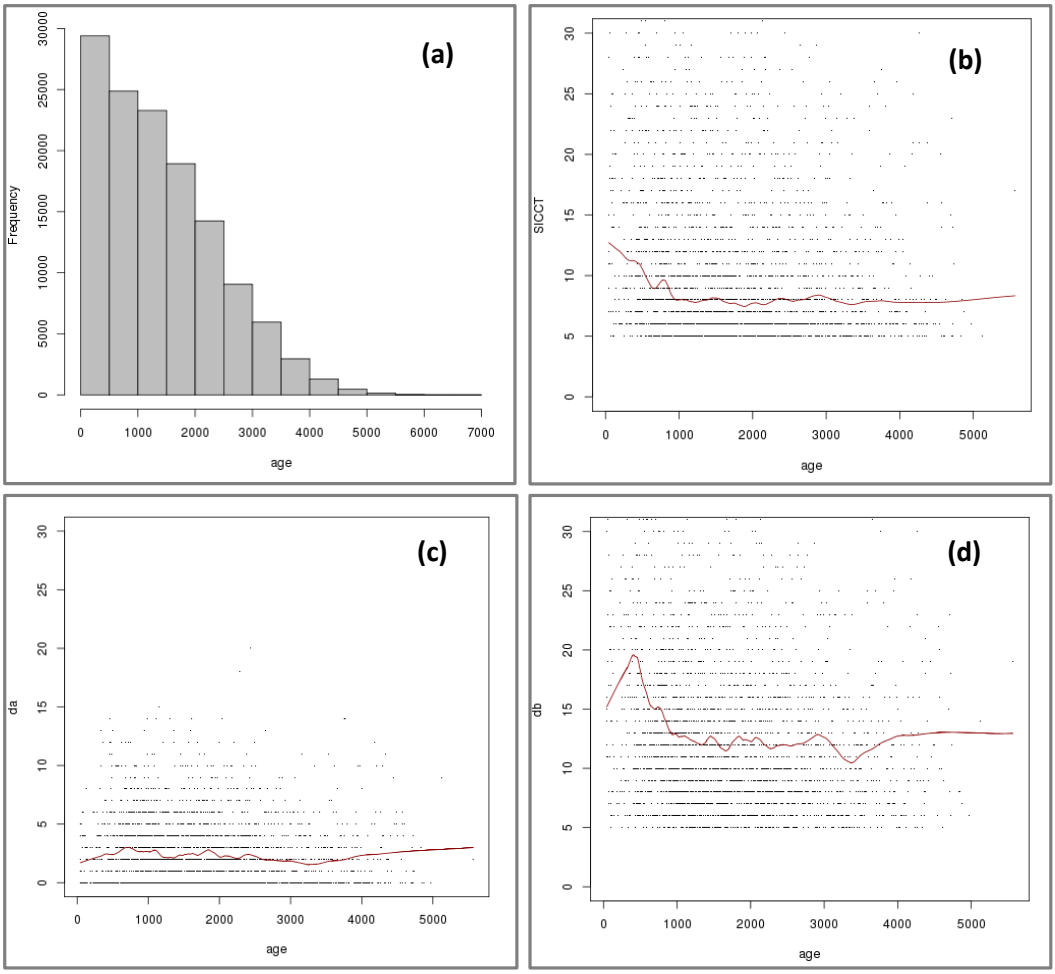


Figure 6. (a) Age distribution in the data; (b) relationship of *SICCT* with age and fitted line using LOESS for the reactors; (c) relationship of *da* with age and fitted line using LOESS for the reactors; (d) relationship of *db* with age and fitted line using LOESS for the reactors; where reactors are defined under the standard interpretation as those with *SICCT>4*.

	Order of polynomial			
	1	**2**	**3**	**4**
SICCT	**591150.4** (k=5)	591847.8 (k=6)	592566.7 (k=7)	593287.7 (k=8)
da	520758.1 (k=5)	**519845.9** (k=6)	519868.1 (k=7)	520493.8 (k=8)

Table 2. AIC results from ANOVA with age fitted as a polynomial of different orders, for the *SICCT* and *da*.

	Group 1	**Group 2**	**Group 3**	**Total**
	age≤ 750	750 < age ≤ 1100	age > 1100	
# records	40788	18438	71373	130599

Table 3. Age ranges and number of records in each of the three different age-groups.

		σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	**(a)**	0.070 (0.016)	5.110 (0.020)	0.014 (0.003)
	(b)	0.064 (0.015)	5.109 (0.020)	0.013 (0.003)
	(c)	0.059 (0.014)	5.102 (0.020)	0.012 (0.003)
	(d)	0.061 (0.015)	5.135 (0.020)	0.012 (0.003)
da	**(a)**	0.138 (0.017)	2.942 (0.012)	0.047 (0.006)
	(b)	0.088 (0.013)	2.942 (0.012)	0.030 (0.004)
db		0.186 (0.029)	6.598 (0.026)	0.028 (0.004)

Table 4a. Variance components and heritability estimates from the preliminary analysis. The age and breakdown are fitted as covariates; test year, test month, test year and test month interaction, and herd are fitted as fixed effect; the composite variable with herd, date of breakdown and lactation group interaction, and the sire are fitted as random effects. For *SICCT* (b) and *da* (a) the age is fitted as a second order polynomial. For *SICCT* (c) and (d), *da* (b), and *db*, age is fitted using a smoothing spline. For *SICCT* (d), *da* (b) and *db* the test year in days has replaced the test year (see Appendix 6.1a).

		σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	**(a)**	0.074 (0.016)	5.001 (0.020)	0.015 (0.003)
	(b)	0.076 (0.016)	5.034 (0.020)	0.015 (0.003)
logSICCT		0.011 (0.001)	0.137 (0.001)	0.081 (0.008)
da		0.121 (0.015)	2.875 (0.012)	0.042 (0.005)
db		0.414 (0.046)	6.472 (0.027)	0.064 (0.007)

Table 4b. Variance components and heritability estimates obtained from preliminary analysis following the additional cleaning of the data (section 6.2.1.2). The age and breakdown are fitted as covariates; test year, test month, test year and test month interaction, the herd, and the season are fitted as fixed effects; Herd by date of breakdown by age group interaction, and sire are fitted as random effects. For *SICCT* (b) the test year in days is fitted in the model. For *da* the age is fitted as a second order polynomial (see Appendix 6.1b).

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.067 (0.015)	4.998 (0.020)	0.013 (0.003)
logSICCT	0.009 (0.001)	0.137 (0.001)	0.064 (0.007)
da	0.122 (0.015)	2.875 (0.012)	0.042 (0.006)
db	0.323 (0.040)	6.442 (0.026)	0.050 (0.006)
SICCT	0.009 (0.004)	2.728 (0.011)	0.003 (0.001)
logSICCT	0.001 (0.000)	0.069 (0.000)	0.015 (0.003)
da	0.107 (0.014)	2.831 (0.011)	0.038 (0.005)
db	0.057 (0.011)	3.484 (0.014)	0.016 (0.003)

Table 5a. Variance components and heritability estimates from the comprehensive analysis. Age was fitted as a polynomial; breakdown was fitted as a covariate; test year by test month interaction and herd were fitted as fixed effects; herd by date of breakdown by age group interaction and sire were fitted as random effects. For the lower part of the table the health status was additionally fitted as a fixed effect (see Appendix 6.2).

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.060 (0.014)	4.993 (0.020)	0.012 (0.003)
logSICCT	0.007 (0.001)	0.136 (0.001)	0.050 (0.006)
da	0.092 (0.013)	2.853 (0.011)	0.032 (0.005)
db	0.218 (0.032)	6.390 (0.026)	0.034 (0.005)
SICCT	0.009 (0.004)	2.727 (0.011)	0.003 (0.001)
logSICCT	0.001 (0.000)	0.068 (0.000)	0.012 (0.003)
da	0.082 (0.012)	2.811 (0.011)	0.029 (0.004)
db	0.037 (0.009)	3.466 (0.014)	0.011 (0.003)

Table 5b. Variance components and heritability estimates obtained from the comprehensive analysis after fitting a smoothing spline for age. For the lower part of the table health status was fitted as a fixed effect (see Appendix 6.2).

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.155 (0.036)	4.178 (0.030)	0.037 (0.009)
logSICCT	0.004 (0.001)	0.055 (0.000)	0.077 (0.013)
da	0.091 (0.019)	1.806 (0.013)	0.050 (0.010)
db	0.166 (0.037)	4.624 (0.033)	0.036 (0.008)
a_1	1.154 (0.070)	1.503 (0.019)	0.768 (0.038)
b_1	1.202 (0.072)	1.540 (0.019)	0.780 (0.038)
SICCT	0.000 (0.005)	2.200 (0.016)	0.000 (0.002)
logSICCT	0.000 (0.000)	0.016 (0.000)	0.011 (0.005)
da	0.086 (0.018)	1.765 (0.013)	0.049 (0.010)
db	0.012 (0.007)	2.050 (0.014)	0.006 (0.003)

Table 6a. ASReml heritability analysis for age-group 1. The models used are the same as described in the comprehensive analysis (see Appendix 6.2). For the lower part of the table health status is additionally fitted as a fixed effect.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.426 (0.108)	5.291 (0.058)	0.081 (0.020)
logSICCT	0.016 (0.004)	0.149 (0.002)	0.110 (0.023)
da	0.114 (0.046)	3.207 (0.035)	0.036 (0.014)
db	1.193 (0.222)	6.991 (0.083)	0.171 (0.031)
a_1	0.143 (0.027)	1.052 (0.012)	0.136 (0.025)
b_1	0.126 (0.026)	1.092 (0.012)	0.116 (0.023)
SICCT	0.016 (0.019)	2.702 (0.029)	0.006 (0.007)
logSICCT	0.002 (0.001)	0.073 (0.001)	0.026 (0.011)
da	0.095 (0.042)	3.151 (0.034)	0.030 (0.013)
db	0.058 (0.035)	3.502 (0.037	0.016 (0.010)

Table 6b. ASReml heritability analysis for age-group 2. The models used are the same as described in the comprehensive analysis (see Appendix 6.2). For the lower part of the table health status is additionally fitted as a fixed effect.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.045 (0.017)	5.266 (0.028)	0.009 (0.003)
logSICCT	0.004 (0.001)	0.170 (0.001)	0.023 (0.005)
da	0.109 (0.019)	3.271 (0.018)	0.033 (0.006)
db	0.161 (0.036)	6.941 (0.037)	0.023 (0.005)
a_1	0.068 (0.008)	0.944 (0.005)	0.072 (0.009)
b_1	0.072 (0.009)	0.956 (0.005)	0.075 (0.009)
SICCT	0.006 (0.005)	2.896 (0.015)	0.002 (0.002)
logSICCT	0.001 (0.000)	0.094 (0.000)	0.008 (0.003)
da	0.106 (0.019)	3.239 (0.017)	0.033 (0.006)
db	0.059 (0.016)	3.980 (0.021)	0.015 (0.004)

Table 6c. ASReml heritability analysis for age-group 3. The models used are the same as described in the comprehensive analysis (see Appendix 6.2). For the lower part of the table health status is additionally fitted as a fixed effect.

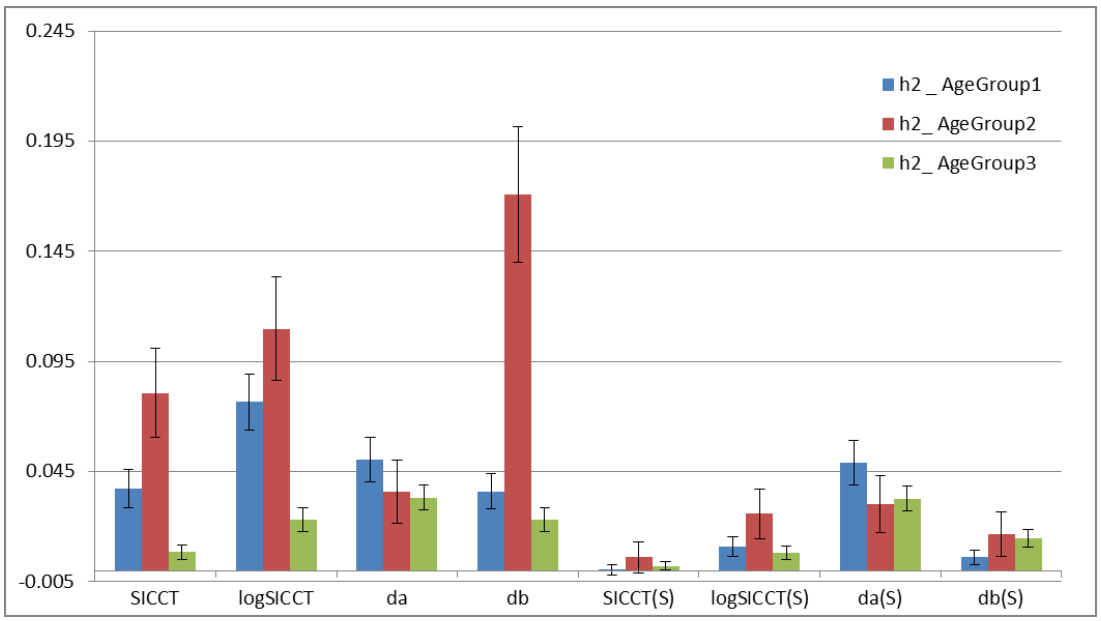


Figure 7. Genomic heritability estimates for each of the three age-groups, corresponding to each of the models tested. The vertical bars represent the ±SE. The (S) denotes that for those models the health status was additionally fitted as a fixed effect.

Cor_G	Group 1	Group 2	Group 3
Group 1	1	0.43 (0.17)	-0.19 (0.25)
Group 2	0.43 (0.17)	1	0.27 (0.24)
Group 3	-0.19 (0.25)	0.27 (0.24)	1

Table 7a. Genetic correlations obtained between the three age-groups from the across-ages bivariate analyses.

	Group 1	Group 2	Group 3
SICCT1 SICCT2	0.036 (0.008)	0.082 (0.020)	-
SICCT2 SICCT3	-	0.083 (0.020)	0.009 (0.003)
SICCT1 SICCT3	0.036 (0.009)	-	0.009 (0.003)

Table 7b. Genomic heritability estimates from the bivariate analyses for each of the age-groups, where SICCT1, SICCT2 and SICCT3 are the SICCT values within age-group 1, age-group 2, and age-group 3 respectively.

Cor_G	EBV 1	EBV 2	EBV 3
EBV 1	1	0.61	0.11
EBV 2	0.61	1	0.41
EBV 3	0.11	0.41	1

Table 8. Approximate genetic correlations calculated from the EBVs for each of the three age-groups after fitting the health status as a fixed effect.

		σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	**(a)**	0.198 (0.030)	5.146 (0.022)	0.039 (0.006)
	(b)	0.196 (0.029)	5.121 (0.022)	0.038 (0.006)
a_1		0.664 (0.028)	1.312 (0.008)	0.506 (0.019)
b_1		0.680 (0.029)	1.339 (0.008)	0.508 (0.019)
da		0.215 (0.023)	3.019 (0.013)	0.071 (0.007)
db		0.911 (0.075)	6.762 (0.031)	0.135 (0.011)
SICCT		0.014 (0.005)	2.772 (0.011)	0.005 (0.002)
a_1		0.664 (0.028)	1.312 (0.008)	0.506 (0.019)
b_1		0.680 (0.029)	1.338 (0.008)	0.508 (0.019)
da		0.181 (0.020)	2.966 (0.013)	0.061 (0.007)
db		0.136 (0.020)	3.572 (0.015)	0.038 (0.006)

Table 9. Heritability estimates and variance components from the first-records analyses. Age is fitted as a covariate, test date and the new breakdown are fitted as fixed effects. Smoothing splines are fitted for age (with *50* knots in (a) and *100* knots in (b)), and for test date. The sire is fitted as a random effect. For the lower part of the table the health status was additionally fitted as a fixed effect (see Appendix 6.3a).

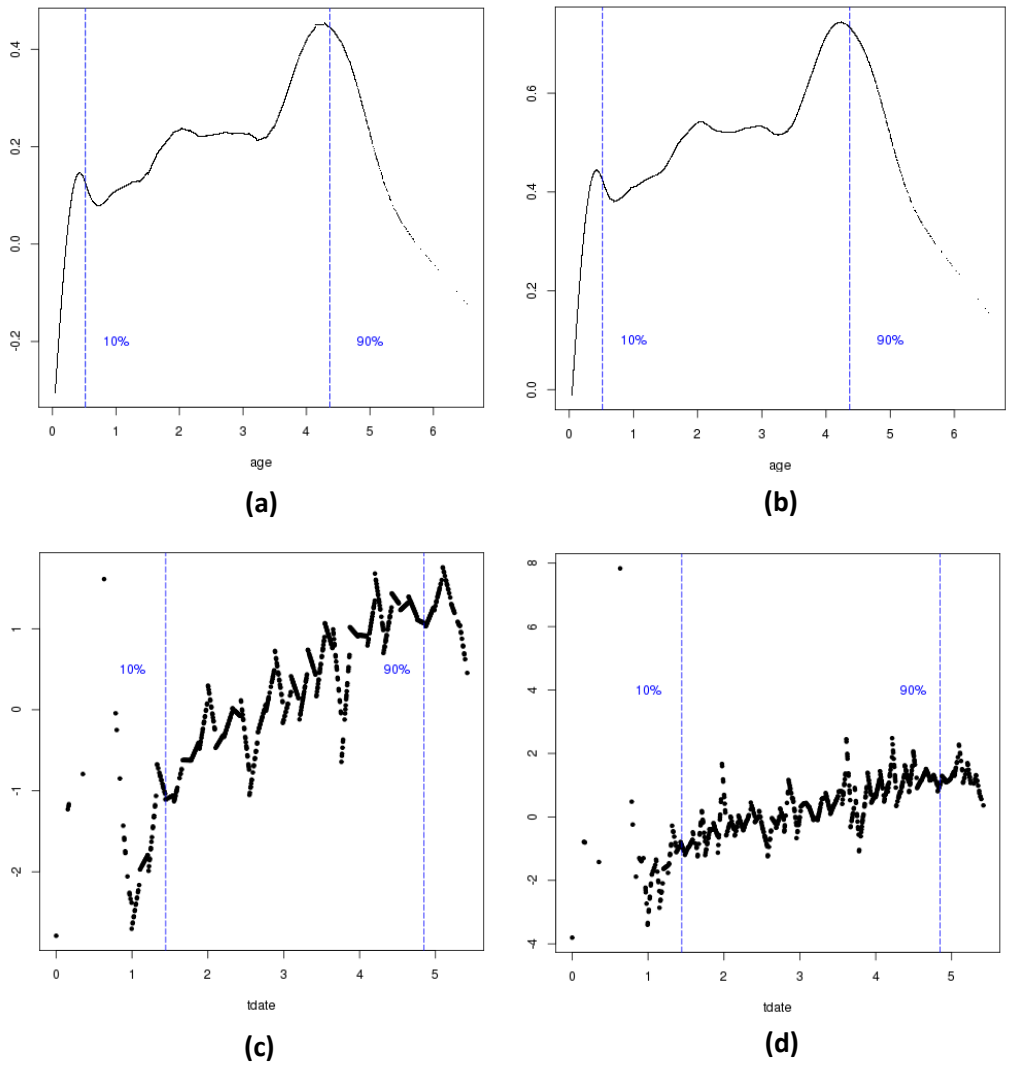


Figure 8. Fitted cubic smoothing splines for age vs. the predicted value from the ASReml analysis with (a) *50*, and (b) *100* knot points. Smoothing spline for test date (*tdate*) with (c) *50*, and (d) *100* knot points.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.038 (0.012)	3.645 (0.017)	0.010 (0.003)
a_1	0.540 (0.026)	1.105 (0.008)	0.489 (0.021)
b_1	0.557 (0.027)	1.128 (0.008)	0.493 (0.021)
da	0.155 (0.020)	2.759 (0.013)	0.056 (0.007)
db	0.257 (0.035)	4.349 (0.021)	0.059 (0.008)

Table 10. Heritability estimates after retaining only the first known test within each new breakdown (see Appendix 6.3b). New breakdown is fitted as a fixed effect, age is fitted using a smoothing spline, and sire is fitted as a random effect.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.029 (0.006)	1.414 (0.007)	0.020 (0.004)
a_1	0.544 (0.026)	1.103 (0.008)	0.493 (0.021)
b_1	0.562 (0.027)	1.128 (0.008)	0.498 (0.022)
da	0.132 (0.018)	2.628 (0.013)	0.050 (0.007)
db	0.060 (0.009)	1.407 (0.007)	0.043 (0.006)

Table 11. Heritability estimates after retaining only the first known test within each new breakdown (see Appendix 6.3b) and after removing the reactors under the standard interpretation.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.032 (0.006)	1.306 (0.006)	0.025 (0.005)
a_1	0.549 (0.027)	1.102 (0.008)	0.498 (0.022)
b_1	0.566 (0.027)	1.126 (0.008)	0.502 (0.022)
da	0.116 (0.017)	2.555 (0.013)	0.045 (0.006)
db	0.043 (0.007)	1.069 (0.005)	0.040 (0.006)

Table 12. Heritability estimates after retaining only the first known test within each new breakdown (see Appendix 6.3b) and after removing the reactors under the severe interpretation.

Trait A	Trait B	Cor_G (SE)	Cor_P (SE)	Trait A h^2 (SE)	Trait B h^2 (SE)	Reg_G	Reg_P
a_1	**b_1**	0.999	0.958	0.493	0.493	1.009	0.967
		(0.000)	(0.000)	(0.021)	(0.021)	(0.005)	(0.001)
a_1	**a_2**	0.910	0.559	0.489	0.215	1.149	1.064
		(0.012)	(0.003)	(0.021)	(0.013)	(0.033)	(0.006)
b_1	**b_2**	0.871	0.477	0.492	0.178	1.168	1.065
		(0.018)	(0.003)	(0.021)	(0.012)	(0.042)	(0.008)
da	**db**	0.901	0.500	0.063	0.053	1.040	0.627
		(0.029)	(0.003)	(0.008)	(0.007)	(0.071)	(0.004)

Table 13. Results from the first-records bivariate analysis after retaining only the first known test within each new breakdown.

Trait A	Trait B	Cor_G (SE)	Cor_P (SE)	Trait A h^2 (SE)	Trait B h^2 (SE)	Reg_G	Reg_P
a_1	b_1	0.962	0.958	0.498	0.498	1.009	0.968
		(0.001)	(0.000)	(0.021)	(0.021)	(0.005)	(0.001)
a_1	a_2	0.923	0.567	0.493	0.220	1.153	1.060
		(0.011)	(0.003)	(0.021)	(0.014)	(0.032)	(0.006)
b_1	b_2	0.960	0.681	0.498	0.284	1.104	1.037
		(0.006)	(0.002)	(0.021)	(0.016)	(0.023)	(0.004)
da	db	0.923	0.682	0.052	0.042	0.607	0.499
		(0.023)	(0.002)	(0.007)	(0.006)	(0.037)	(0.002)

Table 14. Results from the first-records bivariate analysis after retaining only the first known test within each new breakdown and after removing the reactors under the standard interpretation.

Trait A	Trait B	Cor_G (SE)	Cor_P (SE)	Trait A h^2 (SE)	Trait B h^2 (SE)	Reg_G	Reg_P
a_1	b_1	0.999	0.958	0.502	0.502	1.009	0.968
		(0.000)	(0.000)	(0.022)	(0.022)	(0.005)	(0.001)
a_1	a_2	0.931	0.571	0.498	0.217	1.140	1.059
		(0.010)	(0.003)	(0.022)	(0.013)	(0.031)	(0.006)
b_1	b_2	0.970	0.728	0.502	0.307	1.078	1.034
		(0.005)	(0.002)	(0.022)	(0.016)	(0.019)	(0.004)
da	db	0.909	0.702	0.047	0.038	0.532	0.454
		(0.026)	(0.002)	(0.007)	(0.006)	(0.035)	(0.002)

Table 15. Results from the first-records bivariate analysis after retaining only the first known test within each new breakdown and after removing the reactors under the severe interpretation.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.005 (0.013)	2.983 (0.027)	0.002 (0.004)
a_1	0.742 (0.058)	1.156 (0.016)	0.642 (0.043)
b_1	0.817 (0.061)	1.194 (0.017)	0.684 (0.044)
da	0.095 (0.025)	1.964 (0.018)	0.048 (0.013)
db	0.030 (0.016	2.893 (0.026)	0.010 (0.005)

Table 16a. Heritability analysis on the first known test within each new breakdown for age-group 1.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.077 (0.056)	4.145 (0.054)	0.019 (0.014)
a_1	0.089 (0.022)	0.938 (0.013)	0.095 (0.023)
b_1	0.071 (0.020)	0.969 (0.013)	0.073 (0.020)
da	0.039 (0.037)	2.882 (0.038)	0.013 (0.013)
db	0.157 (0.079)	4.728 (0.062)	0.033 (0.017)

Table 16b. Heritability analysis on the first known test within each new breakdown for age-group 2.

	σ_A^2 (SE)	σ_P^2 (SE)	h^2 (SE)
SICCT	0.053 (0.018)	3.660 (0.023)	0.015 (0.005)
a_1	0.074 (0.010)	0.853 (0.006)	0.087 (0.011)
b_1	0.082 (0.010)	0.867 (0.006)	0.094 (0.012)
da	0.096 (0.020)	2.941 (0.019)	0.033 (0.007)
db	0.115 (0.031)	4.652 (0.029)	0.025 (0.007)

Table 16c. Heritability analysis on the first known test within each new breakdown for age-group 3.

Cor_G	Group 1	Group 2	Group 3
Group 1	1	0.999 (B)	-0.643
Group 2	0.999 (B)	1	-0.329
Group 3	-0.643	-0.329	1

Table 17. Genetic correlations for *SICCT* across the different age-groups from the multivariate analysis, where in the parentheses are reported the approximate standard errors of the estimates and B indicates that the estimate was on the boundary.

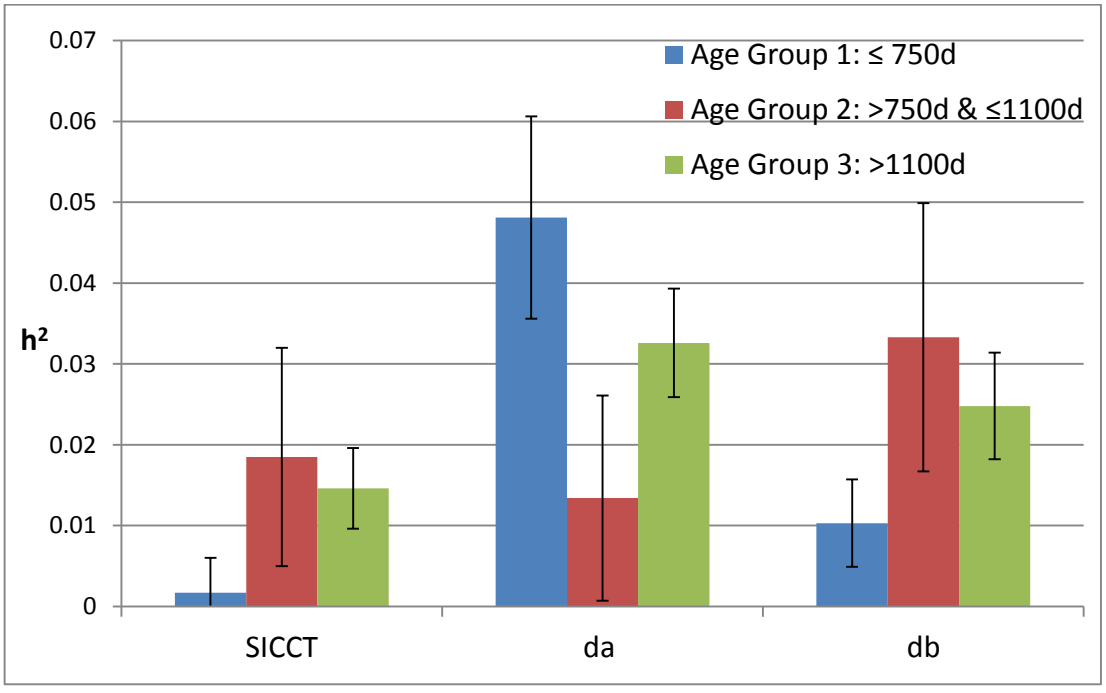


Figure 9. Genomic heritability estimates for the *SICCT*, *da* and *db*, from the analysis on the first known test within each of the new breakdowns, for each of the three age-groups.

Model		Pedigree	σ_A^2	SE	σ_P^2	SE	h_O^2	SE	h_L^2 (p=10%)	h^2_L (p=6%)	SE
LM	SICCT	No	0.086	0.016	5.109	0.020	0.017	0.003	0.049	0.067	-
LM	SICCT	Yes	0.064	0.015	5.109	0.020	0.013	0.003	0.037	0.050	-
LM	QPOS	No	0.003	0.000	0.056	0.000	0.052	0.006	0.153	0.208	-
LM	QPOS	Yes	0.003	0.000	0.056	0.000	0.058	0.007	0.169	0.230	-
LM	QNEG	No	0.000	0.000	0.021	0.000	0.021	0.004	0.060	0.082	-
LM	QNEG	Yes	0.000	0.000	0.021	0.000	0.021	0.004	0.060	0.082	-
TM	QPOS	No	0.439	0.065	3.410	0.016	-	-	0.129		0.019
TM	QPOS	Yes	0.470	0.075	3.417	0.019	-	-	0.137		0.021
TM	QNEG	No	0.569	0.120	3.442	0.030	-	-	0.165		0.033
TM	QNEG	yes	0.583	0.131	3.446	0.033	-	-	0.169		0.036

Table 18. The additive variance, phenotypic variance and heritability estimates on the observed and liability scale for the different models used, using two different values for the assumed population prevalence (Supplementary analysis). In the Table, LM and TM represent the linear and the threshold (logit link function) models used, where for the threshold model σ_A^2 is the additive genetic variance on the liability scale. h_L^2 is calculated using the assumed population prevalence *p=10%* or *p=6%* (section 6.3.5).

	Model		Pedigree	h_O^2	Apparent p	i	h_L^2
1	LM	SICCT	No	0.017	0.023	2.386	0.125
2	LM	SICCT	Yes	0.013	0.023	2.386	0.094
3	LM	QPOS	No	0.052	0.069	1.918	0.192
4	LM	QPOS	Yes	0.058	0.069	1.918	0.212
5	LM	QNEG	No	0.021	0.023	2.386	0.154
6	LM	QNEG	Yes	0.021	0.023	2.386	0.154

Table 19. Heritability estimates on the observed and liability scale for the different models used, using the apparent prevalence specific to the *QPOS* and *QNEG*, for the linear models (Supplementary analysis).

Appendix 6.1

Preliminary analysis

(a) ASReml models used in the preliminary analysis on all 130,626 records (Table 4a).

1) SICCT ~ mu age outbreak tyr tmn tyr.tmn herd !r herd_breakdate_lact sire
2) SICCT ~ mu pol(age,2) outbreak tyr tmn tyr.tmn herd !r herd_breakdate_lact sire
3) da ~ mu pol(age,2) outbreak tyr tmn tyr.tmn herd !r herd_breakdate_lact sire
4) SICCT ~ mu age outbreak tyr tmn tyr.tmn herd !r herd_breakdate_lact sire spl(age)
5) SICCT ~ mu age outbreak tyrd tmn tyrd.tmn herd !r herd_breakdate_lact sire spl(age)
6) da ~ mu age outbreak tyrd tmn !r herd_breakdate_lact sire spl(age)
7) db ~ mu age outbreak tyrd tmn !r herd_breakdate_lact sire spl(age)

(b) ASReml models used in the preliminary analysis on the 130,599 records retained after data cleaning (Table 4b).

1) SICCT ~ mu age outbreak tyr tmn tyr.tmn herd season !r herd_breakdate_group sire
2) SICCT ~ mu age outbreak tyrd tmn tyrd.tmn herd season !r herd_breakdate_group sire
3) logSICCT ~ mu age outbreak tyr tmn tyr.tmn herd season !r herd_breakdate_group sire
4) da ~ mu pol(age,2) outbreak tyr tmn tyr.tmn herd season !r herd_breakdate_group sire
5) db ~ mu age outbreak tyr tmn tyr.tmn herd season !r herd_breakdate_group sire

Appendix 6.2

Comprehensive analysis

(a) ASReml models used in the comprehensive analysis (Table 5a).

(1) SICCT ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire
(2) logSICCT ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire
(3) da ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire
(4) db ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire

(b) ASReml models used in the comprehensive analysis after additionally fitting the health status ("S") as a fixed effect (Table 5a).

(5) SICCT ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire
(6) logSICCT ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire
(7) da ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire
(8) db ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire

(c) ASReml models used in the comprehensive analysis after fitting a smoothing spline for age as a random effect (Table 5b).

1) SICCT ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire spl(age)
2) logSICCT ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire spl(age)
3) da ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire spl(age)
4) db ~ mu pol(age,2) outbreak tyr.tmn herd !r herd_breakdate_group sire spl(age)

5) SICCT ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire spl(age)
6) logSICCT ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire spl(age)
7) da ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire spl(age)
8) db ~ mu pol(age,2) outbreak tyr.tmn herd S !r herd_breakdate_group sire spl(age)

Appendix 6.3

First records analysis

(a) ASReml models used in the first-records analysis (Table 9).

1) SICCT ~ mu age tdate !r sire spl(age,50) spl(tdate,50) !f newoutbreak
1b) SICCT ~ mu age tdate !r sire spl(age,100) spl(tdate,100) !f newoutbreak
2) a1 ~ mu age tdate !r sire spl(age,50) spl(tdate,50) !f newoutbreak
3) b1 ~ mu age tdate !r sire spl(age,50) spl(tdate,50) !f newoutbreak
4) da ~ mu age tdate !r sire spl(age,50) spl(tdate,50) !f newoutbreak
5) db ~ mu age tdate !r sire spl(age,50) spl(tdate,50) !f newoutbreak

6) SICCT ~ mu age tdate S !r sire spl(age,50) spl(tdate,50) !f newoutbreak
7) a1 ~ mu age tdate S !r sire spl(age,50) spl(tdate,50) !f newoutbreak
8) b1 ~ mu age tdate S !r sire spl(age,50) spl(tdate,50) !f newoutbreak
9) da ~ mu age tdate S !r sire spl(age,50) spl(tdate,50) !f newoutbreak
10) db ~ mu age tdate S !r sire spl(age,50) spl(tdate,50) !f newoutbreak

(b) ASReml models used in the first-records analysis when only the first known test within each of the new outbreaks was retained (Tables 10, 11 and 12).

1) SICCT ~ mu age !r sire spl(age) !f newbreakdown
2) a1 ~ mu age !r sire spl(age) !f newbreakdown
3) b1 ~ mu age !r sire spl(age) !f newbreakdown
4) da ~ mu age !r sire spl(age) !f newbreakdown
5) db ~ mu age !r sire spl(age) !f newbreakdown

Chapter 7

General Discussion

7.1 Aims of Thesis and overview of outcomes

The objectives of this Thesis were to study the genetic architecture of resistance to bTB and investigate the genomic control of bTB resistance in cattle, at different levels of the genome, ranging from identifying individual QTLs and groups of SNPs collectively controlling the trait, through exploration of the variance contributed by entire chromosomes, to genomic prediction at the whole genome level. This Thesis aimed to investigate the feasibility of genomic selection for livestock populations that will be more resistant to disease and test the potential of genomic prediction when using high density genotypes and when populations that are only distantly related are combined. Further, this Thesis explored the important issue of an impact of genomic selection for bTB resistance on the response to the diagnostic Single Intradermal Comparative Cervical Test (SICCT) that is traditionally used in the UK for the identification of bTB infected animals.

The results of this PhD have demonstrated that whole genome prediction for disease resistance traits is feasible. Through using dense genetic markers, genomic selection for bTB resistance provided an estimate of prediction accuracy of $r(g, \hat{g}) = 0.33$ and a heritability of $h^2 = 0.23(SE = 0.06)$ in this data. It was shown that by using SNP genotypes, animals that are genetically more resistant to bTB infection can be identified, and selection candidates lacking bTB phenotypes can be selected based on their genotypes. Further, it is possible to focus on the identification of

specific QTLs and the causative mutations. A novel approach is presented that expands the standard GWA analysis to capture underlying non-additive genetic variation, and following this technique a locus displaying heterozygote disadvantage associated with resistance to *M. bovis* was identified on BTA6, with the heterozygotes being significantly more in the cases than in the controls ($p<0.001$).

However, bTB resistance was found to be unlikely to be controlled by a single gene and therefore, my analysis was extended by combining two populations only distantly related in a meta-analysis aiming to provide insight into the genetic architecture of the trait. This analysis revealed that bTB resistance is a moderately polygenic trait, with a few major chromosomes collectively controlling the trait. The results of my meta-analysis have also shown that genomic prediction for bTB is feasible even when only distantly related populations are combined. The prediction accuracy was found to arise from the SNPs capturing Linkage Disequilibrium (LD) between markers and QTLs, as well as additive relationships between animals. Further, Regional Heritability (RH) mapping identified a region on BTA6, putatively associated with bTB resistance. This region was found to contain several plausible candidate genes, including DHX15, which has been shown to be expressed at higher levels in *M. bovis* infected macrophages (Widdison et al. 2008; Mühlbauer L., personal communication, July 8, 2015). High density genotypes were inferred by means of genotype imputation, which is a cost-effective method that facilitates combining populations that have been originally genotyped using different genotyping platforms and allows for the SNPs to be more closely linked to the QTLs. Genotype imputation was found to be successful for dairy cattle, however, the

limiting factor was the number of animals and the trait definitions, rather than the density of genotypes.

Finally, quantitative genetic analysis of field data comprising SICCT values collected during routine bTB herd testing, demonstrated that SICCT has a very low heritability of $h^2=0.0104$ *(SE = 0.0032)*. Thus, genomic selection that will increase bTB resistance is unlikely to genetically alter the SICCT outcome, and any effect on the response to the test is likely to be small.

7.2 Opportunities and implications

7.2.1 Controlling bTB

7.2.1.1 Selection based on phenotypes and genomic selection

Breeding for disease resistance can be considered as a complementary tool for the control of infectious diseases where conventional control strategies have not been successful. Selection based on phenotypes and pedigrees has been effective for endemic diseases where the challenge of infection and exposure to infection can be more predictable, e.g. for nematode infections in sheep (Bishop and Stear 1997; Karlsson et al. 2006), and for mastitis in cattle (Heringstad et al. 2000; http://dairy.ahdb.org.uk/technical-information/breeding-genetics/%C2%A3pli/holstein-reports/#.Vf0pHmeFMeH). However, collecting phenotypes becomes problematic when the population is required to be undergoing an epidemic. BTB is an intermediate situation where although it is an endemic disease, it is not present in all herds, and sporadically, it takes the form of an epidemic. Although selection

based on phenotypes, i.e. EBVs, remains a possibility for bTB, collection of enough phenotypic data to accurately estimate EBVs across the entire population would be challenging and would require the presence of an epidemic. Even if pedigree-recorded herds were affected providing complete and good quality data, analysing this data would only provide results with an application to specific sub-populations, i.e. animals that are more closely related to the herd. For animals that are more distantly related by pedigree to the ones in the epidemic, accuracy of the pedigree-based EBVs would be poor. Moreover, as bTB prevalence declines e.g. at later stages of an eradication programme, identifying cases of bTB based solely on phenotypes will be even more difficult. Thus, selection for bTB resistance using DNA markers becomes a particularly appealing approach as it does not require exposure to infection, and can be potentially very useful even if prediction accuracy is only modest. An important consequence of this is the need for building an adequate training dataset with both phenotypes and genotypes that will allow enhancing the genomic prediction accuracy.

7.2.1.2 On the epidemiology of bTB: reducing the number of secondary cases

Controlling bTB through genomic selection does not necessarily require to drive the beneficial allele to fixation. The basic reproductive number (R_0) defined as the average number of cases generated by one infectious individual, can be informative on whether a control strategy can be successful i.e. if the R_0 can be brought below *1*, the epidemic would die out on its own. Genomic selection for more resistant animals would assist with reducing the number of secondary bTB cases, and

thus, reducing R_0. For bTB in the UK, it has been estimated that the R_0 in cattle is only slightly greater than *1* ($R_0 = 1.07$) (Cox et al. 2005), and thus, even a modest intervention would be sufficient to substantially reduce the risks or severities of bTB breakdowns. Using complementary control strategies such as herd management, control of cattle movements and biosecurity measures can substantially contribute towards bringing an epidemic under control.

This use of complementary control strategies is even more important in the presence of a wildlife reservoir, which is the case for bTB. For heterogeneous populations comprising different host species, such as cattle and badgers in bTB, it is beneficial in the design of control strategies to consider the entire network of infections. Such a framework is provided by the Type-reproduction number (T), i.e. the expected number of cases in one epidemiologically distinct type (e.g. host) generated by one infectious individual of the same type (Roberts et al. 2003; Heesterbeek et al. 2007). For bTB, the secondary cases in cattle generated by one infected cow can be caused either directly or through a sequence of infection events through the badger population. From Brooks-Pollock et al. (2015) it can be derived that the type reproduction number for cattle is $T_c=R_{cc}+[(R_{cb}\ R_{bc})/(1-R_{bb})]$, where R_{bc} and R_{cb} are the number of secondary cases in cattle due to badger-to-cattle transmission and the number of secondary cases in badgers due to cattle-to-badger transmission. Thus, T_c accounts for the amplification effect that might occur after several cycles from entry into the badger population (i.e. R_{cb}) of infection in badgers parameterised by R_{bb}, before infection returns to cattle (i.e. R_{bc}). Similarly, the number of secondary cases in badgers generated by one infected badger can be derived as $T_b=R_{bb}+[(R_{bc}\ R_{cb})/(1-R_{cc})]$. When for all the host species $T<1$, it is also

$R_0<1$ (Heesterbeek et al. 2007). T coincides with R_0 when the population is considered homogeneous and differential transmission from the e.g. the wildlife reservoir, is not taken into account (Heesterbeek et al. 2007; Brooks-Pollock et al. 2015). However, if T in one of the host species is less than 1, it does not necessarily mean that R_0 will also be less than 1. For example, an $R_0>1$ and a major epidemic in the network might occur when T_c in cattle is small but T_b in badgers is large. Further, if both T_c and T_b are greater than 1 but $T_c<T_b<\infty$, then cattle must have the highest reproduction number (i.e. $R_{cc}>R_{bb}$) (Brooks-Pollock et al. 2015), and thus it would be more efficient to control the disease in cattle. Bringing T_c below 1 would result in the control of the epidemic in cattle, however given the data currently available, a wide range of values can be estimated for the R_{cb}, R_{bc}, R_{cc} and R_{bb} (Brooks-Pollock et al. 2015) and thus, explicitly accounting for the individual routes of possible infections that contribute to T_c remains a challenge.

7.2.2 Genetic architecture of bTB resistance: QTL-based selection and genome-wide prediction

The degree of success of Marker Assisted Selection (MAS) based on individual QTLs depends on the proportion of the total genetic variation explained by the QTL. In the present study the hypothesis that it is a single QTL affecting bTB resistance was tested, similarly to what has been observed for the Infectious Pancreatic Necrosis in the Atlantic salmon (Houston et al. 2010). Putative QTLs were identified, however, their individual effects were only small to moderate, and thus, a single QTL is unlikely to explain the total genetic variance seen in bTB resistance. Exploration of the genetic architecture of the trait revealed an

intermediate situation where although there is not a major QTL, it is not infinitesimally polygenic. BTB resistance was found to be moderately polygenic with a handful of QTLs controlling the trait spread across a few chromosomes.

With bTB resistance being more complex than a single gene, genomic selection can assist in improving resistance to disease by means of whole genome prediction. Genomic selection offers a black-box approach to improve health without relying on specific QTLs, and allows genome wide selection (GWS), with a certain accuracy, without requiring specific knowledge on the individual genes and pathways involved. In other words, with genomic selection it is not necessary to know where the QTLs are and what they do.

Nevertheless, genomics provide the opportunity for investigating individual QTLs and causative mutations that are identified, and provide the means for disentangling the exact sources of genetic variation. This can be useful for various scientific purposes, e.g. to provide candidate targets for drug discovery, for exploring the genetic background of response to vaccines, for identifying the individual pathways involved etc. This can be achieved through the use of dense markers that will be more closely linked to the QTLs, or through whole-genome sequence data containing the actual QTLs. For the purposes of EBVs estimation and breeding for bTB resistance in the population as a whole, such information is not required and QTL specific knowledge is not necessary.

Genomic prediction may benefit from taking into account the genetic architecture of the trait. In the present study, results obtained from targeted imputation of genotypes at high density indicated that prediction accuracy might

benefit from being informed by the specific genetic architecture of the trait. Such knowledge can be exploited in genetic evaluations, either through MAS if it was a major QTL segregating, or through Bayesian techniques using genetic architecture as prior information for the estimation of EBVs. It has been previously highlighted in the literature that the genetic architecture of the trait has an impact on the genomic prediction accuracy and incorporating such information in the analyses might be beneficial (Daetwyler et al. 2010; Hayes et al. 2010). For bTB resistance and under a moderately polygenic scenario it would be tempting to utilise such information. However, prediction accuracy still relies on relatedness between animals, while linkage between QTLs and SNPs might differ across different populations. Therefore in conclusion, unless the QTLs have been validated across independent populations, "weighting" for potentially false QTLs could have an undue influence on the predictions by adding noise, and eventually losing selection gain through wasting selection intensity (Sales and Hill 1976).

7.3 Future challenges

7.3.1 Utilising field data in analyses: improving the quality of data

Analyses at present utilise field data opportunistically collected from bTB breakdowns. This data mainly comprise the test values, outcomes of the SICCT. SICCT data currently available is rather noisy, incomplete, and often, with inconsistencies and errors across the records. The need for quality control in bTB test recording has now been recognised by the UK government. Restructuring the recording system, for example by accounting for the relative performance of the individual testers conducting the SICCT (Clegg et al. 2015), and using an electronic

recording system directly informing a national database, can improve the consistency across the records and assist in collecting better quality data. The quality of the data determines the information that can be extracted in analysis, and is of paramount importance for drawing solid conclusions.

Furthermore, SICCT suffers from certain weaknesses that limit its effectiveness in bTB control and reduce its potential in being used as an indicator trait. Mainly due to its imperfect sensitivity, SICCT has not been successful in identifying infected animals at the individual level. SICCT has been used essentially in the same format as it was originally developed and has been only marginally updated (for example in 1975 the mammalian tuberculin was replaced by VLA Weybridge bovine PPD tuberculin, and from 2005 bovine and avian tuberculins from Prionics ID-Lelystad are used). Thus, as will be discussed below in more detail, SICCT could be modified to better reflect the infection status of the modern UK herds. Having a closer look at the test, two important properties can be identified: (a) the *60* days interval in the retesting protocol, and (b) the thresholds for identifying a reactor, an inconclusive reactor, and a non-reactor animal.

The *60* days interval in the modern retesting protocol aims to overcome the suppressive effect of an initial test on the skin reactivity when retesting. Radunz and Lepper (1985) demonstrated that when the animals were retested *4* or *7* days after the initial skin test, there was a suppression of skin reactivity to bovine tuberculin, while at *60* days, reactivity was restored. Doherty et al. (1995) reported a reduced reactivity to a second SICCT within *7* days from the first test, in SICCT-positive *M. bovis* infected cattle. However, in a more recent study it was observed that multiple testing did not affect the sensitivity of the test (Costello et al. 1997), while Thom et al.

(2004) reported that multiple testing before infection did not have an impact on the test outcome post infection. Nevertheless, in the same study it was observed that reactivity to a second test conducted *15* weeks post-challenge with *M. bovis* was reduced compared to reactivity at *7* weeks post-challenge (Thom et al. 2004). Those two tests were ~56 days apart, however, the reduced reactivity could be due to either the first test that took place *56* days earlier, or it could be the effect of time-post-infection, or both. However, limitations in the design of those studies such as the small sample sizes, do not allow for drawing robust conclusions. Additionally, the testing protocols used are different from the modern protocol, and human tuberculin was used instead for bovine tuberculin. Thus, these studies are not directly comparable with the modern SICCT, and they do not reflect accurately its values and limitations. The desensitisation effect remains imperfectly understood and with the view to use vaccines against bTB, it would be crucial to fully disentangle the desensitisation mechanisms which might not be that different from the impacts of a vaccine on the test outcome, potentially compromising SICCT. Hence, it would be beneficial to design studies that will follow the testing protocols currently used in the UK and will have the statistical power to examine the possibility of retesting at shorter intervals. Such studies could be orientated towards the development of a test that its criteria vary with time from the last testing, thus changing SICCT to have a reliable discrimination earlier. This would assist in prompt identification of the infected animals.

The necessity of re-defining SICCT becomes even more apparent if we consider that the thresholds currently used in diagnosing reactors, inconclusive reactors and non-reactors, under the standard or the severe interpretations, might also

be outdated and two possible reasons for that may be tabled. Firstly, although the comparative nature of the test takes into account the presence of *M. avium sbsp avium* and reduces drastically the false positives, the presence of Paratuberculosis (Johne's disease) in the UK herds, caused by *M. avium sbsp paratuberculosis*, and the vaccinations against paratuberculosis, need also to be considered. It is likely that SICCT should be adapted e.g. by modifying the thresholds for classification of animals into diseased and healthy, to take into account the presence of Paratuberculosis which could affect the final value of SICCT. For example, the Paratuberculosis test and the time from the last Paratuberculosis testing could be taken into account as covariates for the SICCT, and that would be anticipated to improve the sensitivity and specificity of the SICCT.

Secondly, although bTB eradication strategies, such as the National Eradication Programme (1950) or the Randomised Badger Culling Trial (1998-2005), might have not been successful in eliminating bTB UK-wide, locally or for a certain period of time they altered bTB prevalence and introduced changes to the infection networks. Moreover, it is possible that *M. bovis* infectiousness might have also changed. This dynamic process of changes that have taken place over time needs to be accounted for by the diagnostic test and therefore SICCT, cannot be static and defined once, but it needs to be re-calibrated. A properly designed diagnostic test would be crucial in bTB control as it would (a) inform more accurately on the progress of the eradication programmes and serve better the purpose of bTB eradication, and (b) be vital in the collection of more informative phenotypes that can be analysed to inform back the design of the eradication programmes.

7.3.2 Improving the prediction accuracy

The results presented in this Thesis demonstrate the feasibility of genomic selection for resistance to infectious diseases. Using genome-wide markers it is possible to predict the risk of infection at the individual level, and select animals more resistant to bTB improving bTB resistance of the population as a whole. The accuracy of prediction presented here, is only a preliminary result. As opposed to selection based on progeny testing and traditional BLUP, the GBLUP accuracy is dynamic, i.e. the value obtained here is not "the prediction accuracy" but it is only the accuracy that could be recovered given the size of the training sets. As the amount of genomic information available increases, this accuracy will also increase, with a theoretical maximum of *1*.

Marker effects are estimated in a training population based on the Linkage Disequilibrium (LD) between markers and QTLs. The results of the meta-analysis presented in Chapter 4 provided evidence that relatedness plays an important role in the proportion of genetic variance captured by the markers. LD between markers and QTLs breaks down over generations due to recombination. Therefore, prediction accuracy over generations declines and marker effects need to be re-calibrated with new phenotypic information. Further, when using markers and not the QTLs themselves, although a marker may be beneficial at present, its effect might change over time due to recombination.

Genomic selection has the potential to assist with overcoming both these limitations. For this purpose, larger datasets with more phenotyped and genotyped animals will be required. Given the scale of the problem of bTB, it would be

reasonable to obtain more samples of cases and controls, for example *5,000* cases/controls, to test if the prediction accuracy will be improved as expected, and then, it would be appropriate to implement genomic selection for bTB control. This scale of ambition is already underway in the RoI as part of their National genomics programme launched in 2014, which aims to collect genomic information on the Irish beef cattle population (http://www.agriculture.gov.ie/press/pressreleases/2014/february/title,73884,en.html). Such information would be of strategic value in deriving genetic evaluations for beef cattle breeding. It would be beneficial to undertake a similar initiative for bTB in the UK, given that the disease has been a long-term problem in the UK cattle industry. In this context, cost-effective methods are available for generating high density genotypes increasingly cheaply, which allows exploitation of all the already available information such as datasets genotyped at lower density. HD genotyping will reduce the impact of recombination across generations as prediction will rely less on relatedness and thus, the prediction accuracy will not decline as quickly. This can be tested once large scale data covering multiple generations is available. Genotyping the entire population at HD and having pedigrees for all the animals will allow for markers more closely linked to the QTLs and less vulnerable to recombination. To take this one step further, progressing to whole genome sequencing will allow prediction based not on the LD between QTLs and markers, but on the QTL itself. Whereas marker effects might change over generations due to recombination, the QTLs remain the same, and thus, there will be no need for re-calibrating the estimated effects.

In the chapters of this Thesis several factors affecting genomic prediction accuracy have been discussed and can be summarised as follows: (i) the genetic

architecture of the trait including the number of the QTLs affecting the trait under study and the distribution of the effects of the QTLs; (ii) the genetic structure of the species, i.e. the effective population size, the number of independent genome segments, and whether there is long-range LD, as the density of markers required to closely capture QTLs will depend on the amount of LD; (iii) the size of the training sets; (iv) the trait heritability; (v) the definition of the phenotypes used and the study design which will have an impact on the estimated marker effects, for example the case/control study design, the imperfect sensitivity and specificity of the diagnostic test, de-regressed EBVs after being weighted and with or without accounting for parent average effects; (vi) the method of estimation of the EBVs, for example BLUP, GBLUP, Bayesian methodology; (vii) relatedness and population structure in the sample, and when proceeding to genomic selection, the genetic distance between the training population and the selection candidates. All these factors determine both the initial prediction accuracy estimates and how genomic selection will progress.

7.3.3 Improving the experimental design

A study can benefit by a carefully chosen experimental design, notably how cases and controls are selected. In Chapter 1, the analysis of a case-control dataset was presented and it was discussed how the case-control design can have an impact on the estimated marker effects, and the genetic (e.g. h^2) and epidemiological (e.g. *Se* and *Sp*) parameters estimated. Analysing field data for endemic livestock diseases or data collected over the course of an epidemic, allow for limited freedom in selecting cases and controls as opposed to challenge experiments under controlled

environments (e.g. in aquaculture). In both cases, there are various approaches that can be followed and will be discussed below.

With regard to the way that controls are selected from the epidemiologist's point of view, controls can be sampled from higher prevalence herds to maximise the probability of exposure. Similarly, controls ideally should be sampled from later stages of an epidemic where the infection pressure and the proportion of animals exposed to the pathogen are high and therefore, those individuals will be more likely to be genuinely resistant (Bishop et al. 2012). In the worst case, these controls will have not been exposed yet or will have been misdiagnosed. The probability of misclassification of a control can be predicted as it depends on the imperfect *Sp* of the diagnostic test used, i.e. *(1-p)(1-Sp)* (Bishop et al. 2010). Nevertheless, in both the cases of unexposed or misdiagnosed controls, these animals will be representative of the herd average animal, given the exposure status of the herd. Although having controls that are true negatives and have been exposed to disease would provide a more powerful contrast, comparing cases to the average can still be useful for extracting information (Bishop et al. 2012). When the controls are a random sample from the whole population, then the experimental design coincides with the "Wellcome Trust design". The "Wellcome Trust design" has been used in human genetic studies, in the context of non-infectious diseases, and the comparison being made is between cases, and controls that represent the population average (Browning and Browning 2008). However, for infectious disease, such a contrast would be even weaker as the comparison of interest is with individuals that have been exposed to the pathogen.

However in statistical analyses, balanced designs, i.e. with one control matched with each case, are more robust. For example, when herds vary considerably with some herds having more controls compared to other herds, if the herd is fitted as a fixed effect, the model will be less efficient and information from herds that do not have both cases and controls will be lost; if the herd is fitted as a random effect some of the inter-herd variation on genetics might be recovered but additional assumptions would have to be made about the validity of the differences between the herds being genetic. A balanced design would allow fitting the herd either as a fixed effect or a random effect and information would be recovered more efficiently in either case. In a genetic analysis, unbalanced designs would not allow consideration in the analysis that different herds might introduce genetic differences which we might want to estimate, i.e. genetic stratification due to different herds and between herds genetic variation. Alternatively, the herd can be considered as a fixed effect and thus removed from genetic interpretation in order to recover information only within herds. However, when for example the controls are selected to originate from higher prevalence herds, then the herd cannot be considered as a fixed effect as it might have an undue influence on the results i.e. results might be potentially biased to a particular subset with specific differences. Either a balanced or an unbalanced design can be used depending on the issues to be addressed in the analysis, and these approaches might not always be reconciled.

7.4 Perspectives for future research and practical considerations

7.4.1 Genomic selection sustainability and multidrug resistance

One of the new challenges arising for disease control is multidrug resistance, i.e. the developed ability of pathogens to be resistant to therapeutic agents, e.g. resistance to antibiotics or anthelminthic resistance (Bennett et al. 2008; Laurenson et al. 2013; Magiorakos et al. 2014). Approaches that will tackle disease while requiring fewer therapeutic treatments need to be developed. Thus, the sustainability of breeding for disease resistance in livestock can be considered with respect to (a) reducing disease without relying on the effectiveness of chemical treatments, (b) pathogen evolution, and (c) improving simultaneously resistance to multiple diseases.

Firstly, genetic selection for disease resistance will improve the average disease resistance of the livestock population. Therefore, it will reduce reliance on treatment-based disease control and assist in dealing with multidrug resistance. Breeding for improved resistance will assist with reducing the in-herd level disease incidence, reducing the likelihood of a breakdown, and reducing the severity of breakdowns. Further, genetic selection is concordant with the modern farming and herd health management strategies which focus on disease prevention, and it can be used as a complementary approach along with other prevention measures e.g. biosecurity.

Secondly, an acknowledged theoretical risk is that some pathogen evolution might occur as a response to selection pressure imposed by genetic selection for disease resistance in the host i.e. host-pathogen coevolution. However, such a response should be compared to pathogen evolution caused by the other disease control strategies that are currently used (e.g. antibiotics, anthelmintic drugs. While the risks deriving from the currently used control strategies are known and drug resistance has developed quickly and dangerously, the potential scale of such a problem due to genetic selection is unknown. Thus at the moment, it remains only a theoretical risk and if it did occur it may only occur over a very long time-scale. Another option is to use knowledge on the genome sequence of the pathogens. However, such approaches are also likely to introduce direct selection pressure on the pathogens, while emerging subtypes of pathogens would require continuous re-calibration. Furthermore, whole-genome based selection for disease resistance in the host, that utilises a large number of QTLs, is likely to be more sustainable compared to single-gene based approaches as pathogens will have to overcome resistance due to multiple loci simultaneously. Alternatively, as will be discussed below in more detail, breeding for disease tolerance is an approach for improving the host's ability to counteract infection without imposing selection pressure on the pathogens.

Lastly, genetic selection for resistance to one disease may be linked to resistance to other diseases. For example, aiming at candidate genes involved in innate immune response has the potential to improve resistance for a wider range of pathogens, while adaptive immune response is more pathogen-specific and prone to increase susceptibility to other pathogens while selecting for resistance to a target-pathogen (Kaiser 2010, p. 15). In a previous study, bovine macrophages were found

to be resistant to *M. bovis* BCG and control its intracellular growth, as well as other intracellular pathogens such as Brucella abortus and Salmonella dublin (Qureshi et al. 1996). Resistance might be possible to be improved simultaneously for pathogens that follow the same route of infection and where the same mechanisms and pathways are involved. Moreover, selection on immunological parameters through e.g identifying genotypes conferring improved immunocompetence and immunoresponsiveness, could be an alternative approach for improving animal health, and models describing these mechanisms have been suggested in the literature (Ask et al. 2007).

7.4.2 Breeding for disease tolerance and reduced infectivity

An alternative approach for disease control is the genetic selection for improved host tolerance, i.e. the ability of the host to limit the damage caused by the pathogen (Kause et al. 2012). Tolerance relates the rate of change in host performance (reproduction and survival (fitness), or production) due to the pathogen, to the pathogen burden within the host (Doeschl-Wilson et al. 2012b), and can be defined either at the group level (Kause 2011) or at the individual level (Doeschl-Wilson et al. 2012b). Improving tolerance presents certain advantages compared to improving resistance, as it is more host-specific and thus likely to be effective against a wider range of pathogens, and it does not place direct selection pressure on the pathogens. However, determining reliable tolerance phenotypes can be challenging and especially when this information needs to be extracted from field data where resistance and tolerance are often confounded (Doeschl-Wilson et al. 2012a and b). Doeschl-Wilson et al. (2012b) introduced the idea of host

performance and pathogen burden trajectories, which have been used to describe the host-pathogen interaction while capturing the dynamic time-dependent interactions between resistance and tolerance over the course of infection (Lough et al. 2014; Lough et al. 2015). Insight into the genetic control of tolerance will allow disentangling the relationship between resistance and tolerance, e.g. if they are antagonistically related, and this knowledge can then be used to inform disease control strategies. Choosing to breed for disease resistance or tolerance will be determined by the aim of a specific disease control programme. For example, breeding for disease resistance would be preferred over tolerance when the aim is to eradicate a disease or when the disease is zoonotic, while breeding for tolerance would be preferred when the disease is endemic with a prevalence approaching unity (Doeschl-Wilson et al. 2012a; Bishop 2012). BTB is a zoonotic disease, and the aim is to eradicate it from the population, therefore, for bTB, breeding for resistance would be preferred over breeding for tolerance.

The presence of genetic variation in infectivity, i.e. differences in the ability of individuals to transmit infection, can be potentially exploited in genetic selection for disease resistance. The impacts of such differences become more apparent if we consider the system of dynamic interactions within a group of animals comprising a mixture of infected and susceptible individuals: the disease status of a member of the group, additionally to its own genetics (own resistance and tolerance), will depend on the infectivity of the other group-members as the latter will influence the levels of exposure to infection for that individual. An extreme example is the case of super-spreaders where only a few individuals are responsible for a large proportion of transmission events. Indirect Genetic Effects (IGE) models have been suggested to

capture the genetic variation in infectivity while accounting for disease dynamics such as the individual's disease status and the time that an individual became infected (Lipschutz-Powell et al. 2012a and b), and the probability of infection was derived, taking into account the force of infection for an individual (i.e. own susceptibility and group members' infectivity) (Lipschutz-Powell et al. 2013). Future studies on the variation in infectivity for specific diseases can potentially inform genomic selection for disease resistance.

7.5 Conclusions

The results presented in this Thesis demonstrate the feasibility of genomic selection for bTB resistance in dairy cattle. Genomic selection holds the promise that given availability of data, it is possible to utilise host genetic variation and breed for resistance to infectious diseases in livestock populations. This approach can be used as a complementary disease control strategy.

Literature cited

Abernethy D. A., Upton P., Higgins I. M., McGrath G., Goodchild A. V., et al. Bovine tuberculosis trends in the UK and the Republic of Ireland, 1995–2010. Veterinary Record. 2013.

AHVLA. Bovine tuberculosis in domestic pets, what this means for you.2014. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/308972/AG-TBYP-01e.pdf.

Allen A. R., Minozzi G., Glass E. J., Skuce R. A., McDowell S. W. J., et al. Bovine tuberculosis: the genetic basis of host susceptibility. Proceedings of the Royal Society B: Biological Sciences. 2010; 277 (1695): 2737-2745. PubMed PMID: PMC2981996.

Altrock P. M., Traulsen A., Reed F. A. Stability Properties of Underdominance in Finite Subdivided Populations. PLoS Comput Biol. 2011; 7 (11): e1002260.

Ameni G., Aseffa A., Engers H., Young D., Gordon S., et al. High Prevalence and Increased Severity of Pathology of Bovine Tuberculosis in Holsteins Compared to Zebu Breeds under Field Cattle Husbandry in Central Ethiopia. Clinical and Vaccine Immunology. 2007; 14 (10): 1356-1361.

Anderson R. M., Trewhella W. Population Dynamics of the Badger (Meles meles) and the Epidemiology of Bovine Tuberculosis (Mycobacterium bovis)1985. p 327-381.

Anderson S. J., Jones R. H. Smoothing splines for longitudinal data. Statistics in Medicine. 1995; 14 (11): 1235-1248.

Andersson L. Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. Genetica. 2009; 136 (2): 341-349.

Anes E., Kuhnel M. P., Bos E., Moniz-Pereira J., Habermann A., et al. Selected lipids activate phagosome actin assembly and maturation resulting in killing of pathogenic mycobacteria. Nat Cell Biol. 2003; 5 (9): 793-802.

Ask B., van der Waaij E. H., Glass E. J., Bishop S. C. Modeling immunocompetence development and immunoresponsiveness to challenge in chicks. Poultry science. 2007; 86 (7): 1336-1350. Epub 2007/06/19. PubMed PMID: 17575180.

Bailey S. S., Crawshaw T. R., Smith N. H., Palgrave C. J. Mycobacterium bovis infection in domestic pigs in Great Britain. Veterinary journal (London, England: 1997). 2013; 198 (2): 391-397. Epub 2013/10/08. PubMed PMID: 24095608.

Barthel R., Piedrahita J. A., McMurray D. N., Payeur J., Baca D., et al. Pathologic findings and association of Mycobacterium bovis infection with the bovine NRAMP1 gene in cattle from herds with naturally occurring tuberculosis. Am J Vet Res. 2000; 61 (9): 1140-1144.

Behr M. A., Wilson M. A., Gill W. P., Salamon H., Schoolnik G. K., et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science. 1999; 284 (5419): 1520-1523. Epub 1999/05/29. PubMed PMID: 10348738.

Bellamy R. Genome-wide approaches to identifying genetic factors in host susceptibility to tuberculosis. Microbes and Infection. 2006; 8 (4): 1119-1123.

Bennett P. M. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. British Journal of Pharmacology. 2008; 153 (Suppl 1): S347-S357. PubMed PMID: PMC2268074.

Bermingham M. L., Bishop S. C., Woolliams J. A., Pong-Wong R., Allen A. R., et al. A genome-wide association study of bovine tuberculosis resistance in the Northern Ireland Holstein-Friesian dairy cattle population BSAS Abstact. 2012.

Bermingham M. L., Bishop S. C., Woolliams J. A., Pong-Wong R., Allen A. R., et al. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. Heredity. 2014; 112 (5): 543-551.

Bermingham M. L., More S. J., Good M., Cromie A. R., Higgins I. M., et al. Genetics of tuberculosis in Irish Holstein-Friesian dairy herds. J Dairy Sci. 2009; 92 (7): 3447-3456.

Binsbergen R. v. Genotype imputation accuracy in Holstein Friesian cattle in case of whole-genome sequence data. eaap Annual Meeting 2013 Nantes. 2013.

Bischof J. M., Chiang A. P., Scheetz T. E., Stone E. M., Casavant T. L., et al. Genome-wide identification of pseudogenes capable of disease-causing gene conversion. Human mutation. 2006; 27 (6): 545-552. Epub 2006/05/04. PubMed PMID: 16671097.

Bishop S. C. A consideration of resistance and tolerance for ruminant nematode infections. Front Genet. 2012; 3 168. Epub 2012/12/19. PubMed PMID: 23248638; PubMed Central PMCID: PMCPmc3522420.

Bishop S. C., Doeschl-Wilson A. B., Woolliams J. A. Uses and implications of field disease data for livestock genomic and genetics studies. Front Genet. 2012; 3 (114): 00114.

Bishop S. C., Stear M. J., Bishop S. C., Stear M. J. Modelling responses to selection for resistance to gastro-intestinal parasites in sheep. Animal science. 1997; 64 469-478.

Bishop S. C., Woolliams J. A. On the genetic interpretation of disease data. PLoS One. 2010; 5 (1): 0008940.

Bishop S. C., Woolliams J. A. Genomics and disease resistance studies in livestock. Livestock Science. 2014; 166 190-198.

Blanco F. C., Bianco M. V., Garbaccio S., Meikle V., Gravisaco M. J., et al. Mycobacterium bovis Deltamce2 double deletion mutant protects cattle against challenge with virulent M. bovis. Tuberculosis (Edinburgh, Scotland). 2013; 93 (3): 363-372. Epub 2013/03/23. PubMed PMID: 23518075.

Boddicker N., Waide E. H., Rowland R. R., Lunney J. K., Garrick D. J., et al. Evidence for a major QTL associated with host response to porcine reproductive and respiratory syndrome virus challenge. J Anim Sci. 2012; 90 (6): 1733-1746. Epub 2011/12/30. PubMed PMID: 22205662.

Boddicker N. J. G., Dorian J.; Reecy, James M.; Rowland, Bob; Lunney, Joan K.; and Dekkers, Jack C. M. Quantitative Trait Locus on Sus scrofa Chromosome 4 Associated with Host Response to Experimental Infection with Porcine Reproductive and Respiratory Syndrome Virus. Animal Industry Report: AS 659, ASL R2823. 2013.

Brendan M. Personal genomes: The case of the missing heritability. Nature. 2008; 456 18-21.

Brooks-Pollock E., Roberts G. O., Keeling M. J. A dynamic model of bovine tuberculosis spread and control in Great Britain. Nature. 2014; 511 (7508): 228-231.

Brooks-Pollock E., Wood J. L. N. Eliminating bovine tuberculosis in cattle and badgers: insight from a dynamic model. 2015.

Brosch R., Pym A. S., Gordon S. V., Cole S. T. The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends in microbiology. 2001; 9 (9): 452-458.

Brotherstone S., White I. M., Coffey M., Downs S. H., Mitchell A. P., et al. Evidence of genetic resistance of cattle to infection with Mycobacterium bovis. J Dairy Sci. 2010; 93 (3): 1234-1242.

Browning B. L., Browning S. R. Haplotypic Analysis of Wellcome Trust Case Control Consortium Data. Human genetics. 2008; 123 (3): 273-280. PubMed PMID: PMC2384233.

Buddle B. M., Pollock J. M., Skinner M. A., Wedlock D. N. Development of vaccines to control bovine tuberculosis in cattle and relationship to vaccine development for other intracellular pathogens. International Journal for Parasitology. 2003; 33 (5–6): 555-566.

Burdick J. T., Chen W.-M., Abecasis G. R., Cheung V. G. In silico method for inferring genotypes in pedigrees. Nat Genet. 2006; 38 (9): 1002-1004.

Cai T. Semi-parametric ROC regression analysis with placement values. Biostatistics. 2004; 5 (1): 45-60.

Calo L. L. Genetic aspects of beef production among Holstein-Friesian cattle pedigree selected for milk and their potential to produce both milk and beef: Cornell Univ.; 1972.

Casati M. Z., Longeri M., Polli M., Ceriotti G., Poli G. BoLA class II polymorphism and immune response to Mycobacterium bovis antigens in vitro. Journal of Animal Breeding and Genetics. 1995; 112 (5-6): 391-400.

Cassidy D. M. L. a. J. P. Guest Editorial, New Perspectives on Bovine Tuberculosis. The Veterinary Journal. 2002; 163 109-110.

Cassidy J. P. The pathogenesis and pathology of bovine tuberculosis with insights from studies of tuberculosis in humans and laboratory animal models. Vet Microbiol. 2006; 112 (2-4): 151-161.

Caws M., Thwaites G., Dunstan S., Hawn T. R., Lan N. T., et al. The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. PLoS Pathog. 2008; 4 (3): 1000034.

Chambers J. C., Elliott P., Zabaneh D., Zhang W., Li Y., et al. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet. 2008; 40 (6): 716-718.

Chambers M. A., Lyashchenko K. P., Greenwald R., Esfandiari J., James E., et al. Evaluation of a rapid serological test for the determination of Mycobacterium bovis infection in badgers (Meles meles) found dead. Clinical and vaccine immunology : CVI. 2010; 17 (3): 408-411. Epub 2010/01/01. PubMed PMID: 20042520; PubMed Central PMCID: PMCPmc2837968.

Chandra N., Kumar D., Rao K. Systems biology of tuberculosis. Tuberculosis (Edinburgh, Scotland). 2011; 91 (5): 487-496.

Charlesworth Brian C. D. Elements of Evolutionary Biology. 2010.

Chatterjee S., Feinstein S. I., Dodia C., Sorokina E., Lien Y.-C., et al. Peroxiredoxin 6 Phosphorylation and Subsequent Phospholipase A2 Activity Are Required for Agonist-mediated Activation of NADPH Oxidase in Mouse Pulmonary Microvascular Endothelium and Alveolar Macrophages. Journal of Biological Chemistry. 2011; 286 (13): 11696-11706.

Chen G. K., Marjoram P., Wall J. D. Fast and flexible simulation of DNA sequence data. Genome Res. 2009; 19 (1): 136-142.

Claridge J., Diggle P., McCann C. M., Mulcahy G., Flynn R., et al. Fasciola hepatica is associated with the failure to detect bovine tuberculosis in dairy cattle. Nat Commun. 2012; 3: 853.

Clegg T. A., Duignan A., More S. J. The relative effectiveness of testers during field surveillance for bovine tuberculosis in unrestricted low-risk herds in Ireland. Prev Vet Med. 2015; 119 (1-2): 85-89. Epub 2015/03/03. PubMed PMID: 25727377.

Cleveland M. A., Hickey J. M., Kinghorn B. P. Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals. BMC Proceedings. 2011; 5 (Suppl 3): S6-S6. PubMed PMID: PMC3103205.

Coad M., Clifford D., Rhodes S., G., Hewinson R., Glyn, Vordermeier H., Martin, et al. Repeat tuberculin skin testing leads to desensitisation in naturally infected tuberculous cattle which is associated with elevated interleukin-10 and decreased interleukin-1 beta responses. Vet Res. 2010; 41 (2): 14.

Cobat A., Barrera L. F., Henao H., Arbelaez P., Abel L., et al. Tuberculin skin test reactivity is dependent on host genetic background in Colombian tuberculosis household contacts. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2012; 54 (7): 968-971. Epub 2012/02/01. PubMed PMID: 22291100; PubMed Central PMCID: PMCPmc3297651.

Cobat A., Gallant C. J., Simkin L., Black G. F., Stanley K., et al. High heritability of antimycobacterial immunity in an area of hyperendemicity for tuberculosis disease. The Journal of infectious diseases. 2010; 201 (1): 15-19. Epub 2009/11/27. PubMed PMID: 19938975.

Cobat A., Gallant C. J., Simkin L., Black G. F., Stanley K., et al. Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. The Journal of experimental medicine. 2009; 206 (12): 2583-2591. Epub 2009/11/11. PubMed PMID: 19901083; PubMed Central PMCID: PMCPmc2806605.

Cobat A., Poirier C., Hoal E., Boland-Auge A., de La Rocque F., et al. Tuberculin skin test negativity is under tight genetic control of chromosomal region 11p14-15 in settings with different tuberculosis endemicities. The Journal of infectious diseases. 2015; 211 (2): 317-321. Epub 2014/08/22. PubMed PMID: 25143445; PubMed Central PMCID: PMCPmc4279780.

Cockett N. E., Jackson S. P., Shay T. L., Farnir F., Berghmans S., et al. Polar Overdominance at the Ovine callipyge Locus. Science. 1996; 273 (5272): 236-238.

Codner G. F., Stear M. J., Reeve R., Matthews L., Ellis S. A. Selective forces shaping diversity in the class I region of the major histocompatibility complex in dairy cattle. Anim Genet. 2012; 43 (3): 239-249. Epub 2012/04/11. PubMed PMID: 22486494.

Coffey M. P., Hickey J., Brotherstone S. Genetic aspects of growth of Holstein-Friesian dairy cows from birth to maturity. J Dairy Sci. 2006; 89 (1): 322-329. Epub 2005/12/17. PubMed PMID: 16357296.

Corbin L., Kranis A., Blott S., Swinburne J., Vaudin M., et al. The utility of low-density genotyping for imputation in the Thoroughbred horse. Genet Sel Evol. 2014; 46 (1): 9. PubMed PMID: doi:10.1186/1297-9686-46-9.

Cosivi O., Grange J. M., Daborn C. J., Raviglione M. C., Fujikura T., et al. Zoonotic tuberculosis due to Mycobacterium bovis in developing countries. Emerg Infect Dis. 1998; 4 (1): 59-70.

Costello E., Egan J. W. A., Quigley F. C., O'Reilly P. F. Performance of the single intradermal comparative tuberculin test in identifying cattle with tuberculous lesions in Irish herds. Veterinary Record. 1997; 141 (9): 222-224.

Cox D. R., Donnelly C. A., Bourne F. J., Gettinby G., McInerney J. P., et al. Simple model for tuberculosis in cattle and badgers. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102 (49): 17588-17593.

Daetwyler H. D., Calus M. P. L., Pong-Wong R., de los Campos G., Hickey J. M. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting and Benchmarking. Genetics. 2012.

Daetwyler H. D., Kemper K. E., van der Werf J. H., Hayes B. J. Components of the accuracy of genomic prediction in a multi-breed sheep population. J Anim Sci. 2012; 90 (10): 3375-3384. Epub 2012/10/06. PubMed PMID: 23038744.

Daetwyler H. D., Pong-Wong R., Villanueva B., Woolliams J. A. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics. 2010; 185 (3): 1021-1031.

Daetwyler H. D. Genome-Wide Evaluation of Populations. PhD Thesis. Animal Breeding and Genomics Centre, Wageningen University, Wageningen, NL 2009.

Daetwyler H. D., Villanueva B., Woolliams J. A. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE. 2008; 3 (10): e3395.

Dardni. Bovine Tuberculosis in Northern Ireland 2013 Annual Report. 2013.

de la Rua-Domenech R., Goodchild A. T., Vordermeier H. M., Hewinson R. G., Christiansen K. H., et al. Ante mortem diagnosis of tuberculosis in cattle: a review of the tuberculin tests, gamma-interferon assay and other ancillary diagnostic techniques. Res Vet Sci. 2006; 81 (2): 190-210.

de Lisle G. W., Collins D. M., Loveday A. S., Young W. A., Julian A. F. A report of tuberculosis in cats in New Zealand, and the examination of strains of Mycobacterium bovis by DNA restriction endonuclease analysis. New Zealand veterinary journal. 1990; 38 (1): 10-13. Epub 1990/04/01. PubMed PMID: 16031566.

de los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Calus M. P. L. Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics. 2012.

Dean G. S., Rhodes S. G., Coad M., Whelan A. O., Cockle P. J., et al. Minimum infective dose of Mycobacterium bovis in cattle. Infection and immunity. 2005; 73 (10): 6467-6471. Epub 2005/09/24. PubMed PMID: 16177318; PubMed Central PMCID: PMCPmc1230957.

DEFRA. Dealing with Bovine TB in your herd 2008. Available from: http://www.rethinkbtb.org/rethink_documents/dealing_with_Tb_in_your_herd_defra.pdf.

DEFRA. A preliminary analysis of existing data to provide evidence of a genetic basis for resistance of cattle to infection with M. bovis and for reactivity to currently used immunological diagnostic tests. - SE3040 2008. Available from: http://randd.defra.gov.uk/Default.aspx?Menu=Menu&Module=More&Location=None&Completed=0&ProjectID=15180.

DEFRA. Extending the quantitative analysis of TB - SE3042 2009. Available from: http://randd.defra.gov.uk/Default.aspx?Menu=Menu&Module=More&Location=None&Completed=2&ProjectID=16720.

DEFRA. Bovine TB Eradication Programme for England 2011. Available from: https://www.gov.uk/government/publications/bovine-tb-eradication-programme-for-england.

DEFRA. Monthly publication of National Statistics on the Incidence of Tuberculosis (TB) in Cattle to end April 2012 for Great Britain 2012. Available from: http://webarchive.nationalarchives.gov.uk/20130123162956/http:/www.defra.gov.uk/statistics/files/defra-stats-foodfarm-landuselivestock-tb-statsnotice-120418.pdf.

DEFRA. Monthly publication of National Statistics on the Incidence of Tuberculosis (TB) in Cattle to end January 2015 for Great Britain 2015. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/433681/bovinetb-statsnotice-13may15.pdf.

Detilleux J., Leroy P. L. Application of a mixed normal mixture model for the estimation of Mastitis-related parameters. J Dairy Sci. 2000; 83 (10): 2341-2349. Epub 2000/10/26. PubMed PMID: 11049078.

Detilleux J., Theron L., Duprez J.-N., Reding E., Humblet M.-F., et al. Structural equation models to estimate risk of infection and tolerance to bovine mastitis. Genet Sel Evol. 2013; 45 (1): 6. PubMed PMID: doi:10.1186/1297-9686-45-6.

Deukhwan Lee D. A. V. Predicting the Accuracy of Breeding Values Using High Density Genome Scans Asian-Aust J Anim Sc. 2011 24 162 - 172.

DFID. Annual Report 2002–2003. Available from: https://www.gov.uk/government/collections/dfid-annual-report-2011-2012.

Divers T., Peek S. Rebhun's Diseases of dairy cattle2008. p 612-614.

Doeschl-Wilson A. B., Bishop S. C., Kyriazakis I., Villanueva B. Novel methods for quantifying individual host response to infectious pathogens for genetic analyses. Frontiers in Genetics. 2012; 3 266. PubMed PMID: PMC3571862.

Doeschl-Wilson A. B., Davidson R., Conington J., Roughsedge T., Hutchings M. R., et al. Implications of Host Genetic Variation on the Risk and Prevalence of Infectious Diseases Transmitted Through the Environment. Genetics. 2011; 188 (3): 683-693. PubMed PMID: PMC3176547.

Doeschl-Wilson A. B., Kyriazakis I., Vincent A., Rothschild M. F., Thacker E., et al. Clinical and pathological responses of pigs from two genetically diverse commercial lines to porcine reproductive and respiratory syndrome virus infection. J Anim Sci. 2009; 87 (5): 1638-1647. Epub 2009/02/03. PubMed PMID: 19181772.

Doeschl-Wilson A. B., Villanueva B., Kyriazakis I. The first step toward genetic selection for host tolerance to infectious pathogens: obtaining the tolerance phenotype through group estimates. Front Genet. 2012; 3 265. Epub 2013/02/16. PubMed PMID: 23412990; PubMed Central PMCID: PMCPmc3571525.

Doherty M. L., Monaghan M. L., Bassett H. F., Quinn P. J. Effect of a recent injection of purified protein derivative on diagnostic tests for tuberculosis in cattle infected with Mycobacterium bovis. Res Vet Sci. 1995; 58 (3): 217-221. Epub 1995/05/01. PubMed PMID: 7659844.

Donnelly C. A., Wei G., Johnston W. T., Cox D. R., Woodroffe R., et al. Impacts of widespread badger culling on cattle tuberculosis: concluding analyses from a large-scale field trial. Int J Infect Dis. 2007; 11 (4): 300-308.

Donnelly C. A., Woodroffe R., Cox D. R., Bourne F. J., Cheeseman C. L., et al. Positive and negative effects of widespread badger culling on tuberculosis in cattle. Nature. 2006; 439 (7078): 843-846.

Donnelly C. A., Woodroffe R., Cox D. R., Bourne J., Gettinby G., et al. Impact of localized badger culling on tuberculosis incidence in British cattle. Nature. 2003; 426 (6968): 834-837.

Dowling D. The thickness of cattle skin. Australian Journal of Agricultural Research. 1955; 6 (5): 776-785.

Dowling D. F. The significance of the thickness of cattle skin. The Journal of Agricultural Science. 1964; 62 (3): 307-311.

Downs S. H., Parry J., Nunez-Garcia J., Abernethy D. A., Broughan J. M., et al. Meta-analysis of diagnostic test performance and modelling of testing strategies for control of bovine tuberculosis in GB. Proceedings of the Society for Veterinary Epidemiology and Preventive Medicine, Leipzig, Germany, 23 Roslin: Society for Veterinary Epidemiology and Preventive Medicine. 2011 139-153.

Driscoll E. E., Hoffman J. I., Green L. E., Medley G. F., Amos W. A Preliminary Study of Genetic Factors That Influence Susceptibility to Bovine Tuberculosis in the British Cattle Herd. PLoS One. 2011; 6 (4): e18806.

Eide D. M., Adnoy T., Klemetsdal G., Nesse L. L., Larsen H. J. Selection for immune response in goats: the antibody response to diphtheria toxoid after 12 years of selection. J Anim Sci. 1991; 69 (10): 3967-3976. Epub 1991/10/01. PubMed PMID: 1778809.

Eide D. M., Adnøy T., Klemetsdal G., Nesse L. L., Larsen H. J. Selection for immune response in goats: the antibody response to diphtheria toxoid after 12 years of selection. Journal of animal science. 1991; 69 (10): 3967-3976.

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. G3 (Bethesda, Md). 2014; 4 (2): 341-347. Epub 2013/12/24. PubMed PMID: 24362310; PubMed Central PMCID: PMCPmc3931567.

Ellis T. M. R., Philips I. R., Lahey T. M. Fortran 90 Programming: Addison Wesley; 1998.

Falconer D. S., Mackay T. F. C. Introduction to Quantitative Genetics. : Longman; 1997.

Fayyazi A., Eichmeyer B., Soruri A., Schweyer S., Herms J., et al. Apoptosis of macrophages and T cells in tuberculosis associated caseous necrosis. The Journal of Pathology. 2000; 191 (4): 417-425.

Ferguson N. M., Donnelly C. A., Anderson R. M. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature. 2001; 413 (6855): 542-548.

Fernando R., Grossman M. Marker assisted selection using best linear unbiased prediction. Genet Sel Evol. 1989; 21 (4): 467 - 477. PubMed PMID: doi:10.1186/1297-9686-21-4-467.

Finlay E. K., Berry D. P., Wickham B., Gormley E. P., Bradley D. G. A Genome Wide Association Scan of Bovine Tuberculosis Susceptibility in Holstein-Friesian Dairy Cattle. PLoS One. 2012; 7 (2): e30545.

Fisher A. B. Peroxiredoxin 6: a bifunctional enzyme with glutathione peroxidase and phospholipase A(2) activities. Antioxid Redox Signal. 2011; 15 (3): 831-844.

Francis J., Seiler R., Wilkie I., O'Boyle D., Lumsden M., et al. The sensitivity and specificity of various tuberculin tests using bovine PPD and other tuberculins. Veterinary Record. 1978; 103 (19): 420-425.

Fridley B. L., Jenkins G., Deyo-Svendsen M. E., Hebbring S., Freimuth R. Utilizing genotype imputation for the augmentation of sequence data. PLoS One. 2010; 5 (6): e11018. Epub 2010/06/15. PubMed PMID: 20543988; PubMed Central PMCID: PMCPmc2882389.

Fujii J., Ikeda Y. Advances in our understanding of peroxiredoxin, a multifunctional, mammalian redox protein. Redox Report. 2002; 7 (3): 123-130.

Gallagher M. J., Higgins I. M., Clegg T. A., Williams D. H., More S. J. Comparison of bovine tuberculosis recurrence in Irish herds between 1998 and 2008. Prev Vet Med. 2013; 111 (3-4): 237-244. Epub 2013/06/12. PubMed PMID: 23746572.

Gammack D., Doering C. R., Kirschner D. E. Macrophage response to Mycobacterium tuberculosis infection. Journal of mathematical biology. 2004; 48 (2): 218-242. Epub 2004/01/28. PubMed PMID: 14745511.

Gao X., Starmer J., Martin E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol. 2008; 32 (4): 361-369. Epub 2008/02/14. PubMed PMID: 18271029.

Garnier T., Eiglmeier K., Camus J.-C., Medina N., Mansoor H., et al. The complete genome sequence of Mycobacterium bovis. Proceedings of the National Academy of Sciences. 2003; 100 (13): 7877-7882.

Garrick D. J., Taylor J. F., Fernando R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009; 41 55. Epub 2010/01/02. PubMed PMID: 20043827; PubMed Central PMCID: PMCPmc2817680.

Gianola D., Fernando R. L., Stella A. Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics. 2006; 173 (3): 1761-1776.

Gilbert M., Mitchell A., Bourn D., Mawdsley J., Clifton-Hadley R., et al. Cattle movements and bovine tuberculosis in Great Britain. Nature. 2005; 435 (7041): 491-496.

Gilmour A., Gogel, B., Cullis, B. & Thompson. R. ASReml User Guide Release 3.0. VSN International Ltd. Hemel Hempstead, HP1 1ES, UK. 2009.

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

Goddard M. E., Hayes B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009; 10 (6): 381-391.

Gonda M. G., Chang Y. M., Shook G. E., Collins M. T., Kirkpatrick B. W. Genetic variation of Mycobacterium avium ssp. paratuberculosis infection in US Holsteins. J Dairy Sci. 2006; 89 (5): 1804-1812.

Gormley E., Doyle M. B., McGill K., Costello E., Good M., et al. The effect of the tuberculin test and the consequences of a delay in blood culture on the sensitivity of a gamma-interferon assay for the detection of Mycobacterium bovis infection in cattle. Vet Immunol Immunopathol. 2004; 102 (4): 413-420.

Groen A. F., Vos H. Genetic parameters for body weight and growth in Dutch Black and White replacement stock. Livestock Production Science. 1995; 41 (3): 201-206.

Gros P., Skamene E., Forget A. Genetic control of natural resistance to Mycobacterium bovis (BCG) in mice. J Immunol. 1981; 127 (6): 2417-2421.

Gutiérrez J. P., Goyache F. A note on ENDOG: a computer program for analysing pedigree information. Journal of Animal Breeding and Genetics. 2005; 122 (3): 172-176.

Habier D., Fernando R. L., Dekkers J. C. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177 (4): 2389-2397.

Habier D., Fernando R. L., Kizilkaya K., Garrick D. J. Extension of the bayesian alphabet for genomic selection. BMC bioinformatics. 2011; 12 186.

Hansen M., Lund M. S., Sorensen M. K., Christensen L. G. Genetic parameters of dairy character, protein yield, clinical mastitis, and other diseases in the Danish Holstein cattle. J Dairy Sci. 2002; 85 (2): 445-452.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer; 2003.

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009; 92 (2): 433-443.

Hayes B. J., Pryce J., Chamberlain A. J., Bowman P. J., Goddard M. E. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet. 2010; 6 (9): e1001139.

Hayes B. J., Visscher P. M., Goddard M. E. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res. 2009; 91 (1): 47-60.

Hayes B. J., Visscher P. M., McPartlan H. C., Goddard M. E. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 2003; 13 (4): 635-643. Epub 2003/03/26. PubMed PMID: 12654718; PubMed Central PMCID: PMCPmc430161.

Heesterbeek J. A., Roberts M. G. The type-reproduction number T in models for infectious disease control. Mathematical biosciences. 2007; 206 (1): 3-10. Epub 2006/03/15. PubMed PMID: 16529777.

Hemani G., Knott S., Haley C. An Evolutionary Perspective on Epistasis and the Missing Heritability. PLoS Genet. 2013; 9 (2): e1003295.

Henderson C. R. Comparison of Alternative Sire Evaluation Methods. Journal of Animal Science. 1975; 41 (3): 760-770.

Heringstad B., Chang Y. M., Gianola D., Osteras O. Short communication: Genetic analysis of respiratory disease in Norwegian Red calves. J Dairy Sci. 2008; 91 (1): 367-370.

Heringstad B., Klemetsdal G., Ruane J. Selection for mastitis resistance in dairy cattle: a review with focus on the situation in the Nordic countries. Livestock Production Science. 2000; 64 (2): 95-106.

Hewinson R. G., Vordermeier H. M., Smith N. H., Gordon S. V. Recent advances in our knowledge of Mycobacterium bovis: A feeling for the organism. Vet Microbiol. 2006; 112 (2-4): 127–139.

Hickey J., Kinghorn B. P., Cleveland M. A., Tier B., van der Werf J. H. J., et al. Recursive long range phasing and long haplotype library imputation: Building a global haplotype library for Holstein cattle. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production (WCGALP)2010. p. 0934.

Hickey J., Kinghorn B. P., Tier B., van der Werf J. H. J., Hickey J., et al. Phasing of SNP data by combined recursive long range phasing and long range haplotype imputation. Proceedings of the 18th Conference of the Association for the Advancement of Animal Breeding and Genetics [AAABG] 2009. 182009. p. 72-75.

Hickey J. M., Gorjanc G. Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods. G3: Genes|Genomes|Genetics. 2012; 2 (4): 425-427.

Hickey J. M., Kinghorn B. P., Tier B., Wilson J. F., Dunstan N., et al. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol. 2011; 43 (12): 1297-9686.

Hill A. V. S. Aspects of genetic susceptibility to human infectious diseases. Annual review of genetics. Annual Review of Genetics. 40. Palo Alto: Annual Reviews; 2006. p. 469-486.

Hinger M., Brandt H., Erhardt G. Heritability estimates for antibody response to Mycobacterium avium subspecies paratuberculosis in German Holstein cattle. J Dairy Sci. 2008; 91 (8): 3237-3244. Epub 2008/07/25. PubMed PMID: 18650301.

Hirotsune S., Yoshida N., Chen A., Garrett L., Sugiyama F., et al. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature. 2003; 423 (6935): 91-96.

Hope J. C., Thom M. L., Villarreal-Ramos B., Vordermeier H. M., Hewinson R. G., et al. Exposure to Mycobacterium avium induces low-level protection from Mycobacterium bovis infection but compromises diagnosis of disease in cattle. Clinical and experimental immunology. 2005; 141 (3): 432-439. Epub 2005/07/28. PubMed PMID: 16045732; PubMed Central PMCID: PMCPmc1809462.

Hou Y., Liu G., Bickhart D., Cardone M., Wang K., et al. Genomic characteristics of cattle copy number variations. BMC Genomics. 2011; 12 (1): 127. PubMed PMID: doi:10.1186/1471-2164-12-127.

Hou Y., Liu G. E., Bickhart D. M., Matukumalli L. K., Li C., et al. Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. Functional & integrative genomics. 2012; 12 (1): 81-92. Epub 2011/09/20. PubMed PMID: 21928070.

Houston R. D., Haley C. S., Hamilton A., Guy D. R., Mota-Velasco J. C., et al. The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. Heredity (Edinb). 2010; 105 (3): 318-327. Epub 2009/11/26. PubMed PMID: 19935825.

Hoze C., Fouilloux M.-N., Venot E., Guillaume F., Dassonneville R., et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013; 45 (1): 33. PubMed PMID: doi:10.1186/1297-9686-45-33.

Huqun, Izumi S., Miyazawa H., Ishii K., Uchiyama B., et al. Mutations in the SLC34A2 gene are associated with pulmonary alveolar microlithiasis. American journal of respiratory and critical care medicine. 2007; 175 (3): 263-268. Epub 2006/11/11. PubMed PMID: 17095743.

Ilska J., Kranis A., Woolliams J. A. The Effect of Training Population Size and Chip Density on Accuracy and Bias of Genomic Predictions in Broiler Chickens. Proceedings, 10th World Congress of Genetics Applied to Livestock Production. 2014.

Institute of Environmental Science & Research Limited P. C., Dr Rob Lake, Dr Andrew Hudson, . Risk Profile: Mycobacterium Bovis in red meat, Prepared as part of a New Zealand Food Safety Authority contract for scientific services. 2006.

Janes H., Pepe M. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. Biostatistics. 2006; 7 (3): 456-468.

Janssens A. C., Aulchenko Y. S., Elefante S., Borsboom G. J., Steyerberg E. W., et al. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med. 2006; 8 (7): 395-400.

Janssens A. C., Moonesinghe R., Yang Q., Steyerberg E. W., van Duijn C. M., et al. The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. Genet Med. 2007; 9 (8): 528-535.

Jenkins H. E., Woodroffe R., Donnelly C. A. The Duration of the Effects of Repeated Widespread Badger Culling on Cattle Tuberculosis Following the Cessation of Culling. PLoS One. 2010; 5 (2): e9090.

Jepson A., Fowler A., Banya W., Singh M., Bennett S., et al. Genetic regulation of acquired immune responses to antigens of Mycobacterium tuberculosis: a study of twins in West Africa. Infection and immunity. 2001; 69 (6): 3989-3994. Epub 2001/05/12. PubMed PMID: 11349068; PubMed Central PMCID: PMCPmc98461.

Jia Y., Jannink J. L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 2012; 192 (4): 1513-1522.

Johansson M. W., Holmblad T., Thornqvist P. O., Cammarata M., Parrinello N., et al. A cell-surface superoxide dismutase is a binding protein for peroxinectin, a cell-adhesive peroxidase in crayfish. Journal of cell science. 1999; 112 (Pt 6) 917-925. Epub 1999/02/26. PubMed PMID: 10036241.

Jørgensen C. B., Cirera S., Anderson S. I., Archibald A. L., Raudsepp T., et al. Linkage and comparative mapping of the locus controlling susceptibility towards E. coli F4ab/ac diarrhoea in pigs. Cytogenetic and Genome Research. 2003; 102 (1-4): 157-162.

Jörgensen C. B., Cirera S., Archibald A., Andersson L., Fredholm M., et al. Porcine polymorphisms and methods for detecting them. Google Patents; 2010.

Kadri N. K., Sahana G., Charlier C., Iso-Touru T., Guldbrandtsen B., et al. A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock. PLoS Genet. 2014; 10 (1): e1004049.

Kandouz M., Bier A., Carystinos G. D., Alaoui-Jamali M. A., Batist G. Connexin43 pseudogene is expressed in tumor cells and inhibits growth. Oncogene. 2004; 23 (27): 4763-4770.

Karlsson L. J. E., Greeff J. C. Selection response in fecal worm egg counts in the Rylington Merino parasite resistant flock. Australian Journal of Experimental Agriculture. 2006; 46 (7): 809-811.

Kause A. Genetic analysis of tolerance to infections using random regressions: a simulation study. Genetics research. 2011; 93 (4): 291-302. Epub 2011/07/20. PubMed PMID: 21767462.

Kearney J. F., Wall E., Villanueva B., Coffey M. P. Inbreeding trends and application of optimized selection in the UK Holstein population. J Dairy Sci. 2004; 87 (10): 3503-3509.

Kerr W. R., Lamont H. G., Mc G. J. Studies on tuberculin sensitivity in the bovine; the Stormont test. The Veterinary record. 1946; 58 (42): 451-454. Epub 1946/10/19. PubMed PMID: 20275312.

Kerr W. R., McGirr J. L., Robertson M. Specific and Non-Specific Desensitisation of the Skin in Trichomonas Sensitive Bovines. Journal of Comparative Pathology and Therapeutics. 1949; 59 (0): 133-154.

Khoury M. J., Newill C. A., Chase G. A. Epidemiologic evaluation of screening for risk factors: application to genetic screening. Am J Public Health. 1985; 75 (10): 1204-1208.

Kim E. S., Kirkpatrick B. W. Linkage disequilibrium in the North American Holstein population. Anim Genet. 2009; 40 (3): 279-288.

Kinghorn B., Hickey J., Van Der Werf J. H. A Recursive Algorithm For Long Range Phasing Of SNP Genotype.

Kinghorn B. P., Hickey J., van der Werf J. H. J. Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production (WCGALP) 2010. p. 0036.

Koets A. P., Adugna G., Janss L. L., van Weering H. J., Kalis C. H., et al. Genetic variation of susceptibility to Mycobacterium avium subsp. paratuberculosis infection in dairy cattle. J Dairy Sci. 2000; 83 (11): 2702-2708.

Korneev S. A., Park J. H., O'Shea M. Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. The Journal of neuroscience : the official journal of the Society for Neuroscience. 1999; 19 (18): 7711-7720. Epub 1999/09/10. PubMed PMID: 10479675.

Koul A., Herget T., Klebl B., Ullrich A. Interplay between mycobacteria and host signalling pathways. Nat Rev Micro. 2004; 2 (3): 189-202.

Kumar D., Nath L., Kamal M. A., Varshney A., Jain A., et al. Genome-wide analysis of the host intracellular network that regulates survival of Mycobacterium tuberculosis. Cell. 2010; 140 (5): 731-743.

Laura C. The Application of Genomic Technologies to the Horse. PhD thesis, University of Edinburgh. 2012.

Laurenson Y. C., Bishop S. C., Forbes A. B., Kyriazakis I. Modelling the short- and long-term impacts of drenching frequency and targeted selective treatment on the performance of grazing lambs and the emergence of anthelmintic resistance. Parasitology. 2013; 140 (6): 780-791. Epub 2013/02/02. PubMed PMID: 23369535.

Lee Sang H., Wray Naomi R., Goddard Michael E., Visscher Peter M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. The American Journal of Human Genetics. 2011; 88 (3): 294-305.

Lee T. H., Yu S. L., Kim S. U., Kim Y. M., Choi I., et al. Characterization of the murine gene encoding 1-Cys peroxiredoxin and identification of highly homologous genes. Gene. 1999; 234 (2): 337-344.

Legarra A., Robert-Granié C., Manfredi E., Elsen J.-M. Performance of Genomic Selection in Mice. Genetics. 2008; 180 (1): 611-618.

Lesslie I., Herbert C., Barnett D. Comparison of the specificity of human and bovine tuberculin PPD for testing cattle. 2. South-eastern England. Veterinary Record. 1975; 96 (15): 335-338.

Leutenegger A.-L., Prum B., Génin E., Verny C., Lemainque A., et al. Estimation of the Inbreeding Coefficient through Use of Genomic Data. American journal of human genetics. 2003; 73 (3): 516-523.

Li X., Yang Y., Zhou F., Zhang Y., Lu H., et al. SLC11A1 (NRAMP1) Polymorphisms and Tuberculosis Susceptibility: Updated Systematic Review and Meta-Analysis. PLoS ONE. 2011; 6 (1): e15831.

Li Y., Willer C., Sanna S., Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10: 387-406.

Li Y., Willer C. J., Ding J., Scheet P., Abecasis G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology. 2010; 34 (8): 816-834.

Lillehammer M., Odegard J., Madsen P., Gjerde B., Refstie T., et al. Survival, growth and sexual maturation in Atlantic salmon exposed to infectious pancreatic necrosis: a multi-variate mixture model approach. Genet Sel Evol. 2013; 45 (1): 8. PubMed PMID: doi:10.1186/1297-9686-45-8.

Lin C. Y., McAllister A. J., Lee A. J. Multitrait Estimation of Relationships of First-Lactation Yields to Body Weight Changes in Holstein Heifers1. Journal of Dairy Science. 1985; 68 (11): 2954-2963.

Lipschutz-Powell D., Woolliams J. A., Bijma P., Doeschl-Wilson A. B. Indirect Genetic Effects and the Spread of Infectious Disease: Are We Capturing the Full Heritable Variation Underlying Disease Prevalence? PLoS ONE. 2012; 7 (6): e39551.

Lipschutz-Powell D., Woolliams J. A., Bijma P., Pong-Wong R., Bermingham M. L., et al. Bias, accuracy, and impact of indirect genetic effects in infectious diseases. Front Genet. 2012; 3 215. Epub 2012/10/25. PubMed PMID: 23093950; PubMed Central PMCID: PMCPmc3477629.

Lipschutz-Powell D., Woolliams J. A., Doeschl-Wilson A. B. A unifying theory for genetic epidemiological analysis of binary disease data. Genet Sel Evol. 2014; 46 15. Epub 2014/02/21. PubMed PMID: 24552188; PubMed Central PMCID: PMCPmc3996085.

Lough G. K. I., Forni S., Doeschl-Wilson A. Dynamic and Genetic Signatures of Resistance and Tolerance of pigs to PRRS. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada, Asas. 2014.

Lough G. K. I., Bergmann S., Lengeling A. and Doeschl-Wilson A.B. Health trajectories reveal the dynamic contributions of host genetic resistance and tolerance to infection outcome. Proc R Soc B Under Review. 2015.

Lowrie D. B., Tascon R. E., Bonato V. L. D., Lima V. M. F., Faccioli L. H., et al. Therapy of tuberculosis in mice by DNA vaccination. Nature. 1999; 400 (6741): 269-271.

Lu H., Lu N., Weng L., Yuan B., Liu Y.-j., et al. DHX15 Senses Double-Stranded RNA in Myeloid Dendritic Cells. The Journal of Immunology. 2014.

Luan T., Woolliams J. A., Lien S., Kent M., Svendsen M., et al. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics. 2009; 183 (3): 1119-1126.

MacHugh D. E., Gormley E., Park S. D., Browne J. A., Taraktsoglou M., et al. Gene expression profiling of the host response to Mycobacterium bovis infection in cattle. Transbound Emerg Dis. 2009; 56 (6-7): 204-214.

MacKenzie K., Bishop S. C. Developing stochastic epidemiological models to quantify the dynamics of infectious diseases in domestic livestock. J Anim Sci. 2001; 79 (8): 2047-2056.

MacKenzie K., Bishop S. C. Utilizing stochastic genetic epidemiological models to quantify the impact of selection for resistance to infectious diseases in domestic livestock. J Anim Sci. 2001; 79 (8): 2057-2065.

Mackintosh C. G., Qureshi T., Waldrup K., Labes R. E., Dodds K. G., et al. Genetic Resistance to Experimental Infection with Mycobacterium bovis in Red Deer (Cervus elaphus). Infection and immunity. 2000; 68 (3): 1620-1625.

Magiorakos A. P., Srinivasan A., Carey R. B., Carmeli Y., Falagas M. E., et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clinical Microbiology and Infection. 2012; 18 (3): 268-281.

Marufu M. C., Chimonyo M., Mans B. J., Dzama K. Cutaneous hypersensitivity responses to Rhipicephalus tick larval antigens in pre-sensitized cattle. Ticks and tick-borne diseases. 2013; 4 (4): 311-316. Epub 2013/03/05. PubMed PMID: 23453577.

Mc Parland S., Kearney J. F., Rath M., Berry D. P. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. J Anim Sci. 2007; 85 (2): 322-331.

McDonald R. A. Animal health: How to control bovine tuberculosis. Nature. 2014; 511 (7508): 158-159.

McQuillan R., Eklund N., Pirastu N., Kuningas M., McEvoy B. P., et al. Evidence of Inbreeding Depression on Human Height. PLoS Genet. 2012; 8 (7): e1002655.

Meade K. G., Gormley E., Park S. D. E., Fitzsimons T., Rosa G. J. M., et al. Gene expression profiling of peripheral blood mononuclear cells (PBMC) from Mycobacterium bovis infected cattle after in vitro antigenic stimulation with purified protein derivative of tuberculin (PPD). Veterinary Immunology and Immunopathology. 2006; 113 (1–2): 73-89.

Metz C. E. Basic principles of ROC analysis. Seminars in nuclear medicine. 1978; 8 (4): 283-298. Epub 1978/10/01. PubMed PMID: 112681.

Meuwissen T. H. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol. 2009; 41 35. Epub 2009/06/13. PubMed PMID: 19519896; PubMed Central PMCID: PMCPmc2708128.

Meuwissen T. H., Goddard M. E. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet Sel Evol. 2004; 36 (3): 261-279.

Meuwissen T. H., Hayes B. J., Goddard M. E. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157 (4): 1819-1829.

Meuwissen T. H. E., Goddard M. E. The use of marker haplotypes in animal breeding schemes. Genet Sel Evol. 1996; 28 (2): 1-16.

Minozzi G., Buggiotti L., Stella A., Strozzi F., Luini M., et al. Genetic Loci Involved in Antibody Response to Mycobacterium avium ssp. paratuberculosis in Cattle. PLoS ONE. 2010; 5 (6): e11117.

Monaghan M. L., Doherty M. L., Collins J. D., Kazda J. F., Quinn P. J. The tuberculin test. Vet Microbiol. 1994; 40 (1-2): 111-124. Epub 1994/05/01. PubMed PMID: 8073619.

Monies B., Jahans K., de la Rua R. Bovine tuberculosis in cats. Veterinary Record. 2006; 158 (7): 245-246.

Morrison W. I., Bourne F. J., Cox D. R., Donnelly C. A., Gettinby G., et al. Pathogenesis and diagnosis of infections with Mycobacterium bovis in cattle. Independent Scientific Group on Cattle TB. The Veterinary record. 2000; 146 (9): 236-242. Epub 2000/03/29. PubMed PMID: 10737292.

Mortensen H., Nielsen S. S., Berg P. Genetic variation and heritability of the antibody response to Mycobacterium avium subspecies paratuberculosis in Danish Holstein cows. J Dairy Sci. 2004; 87 (7): 2108-2113. Epub 2004/08/26. PubMed PMID: 15328223.

Moser G., Lee S. H., Hayes B. J., Goddard M. E., Wray N. R., et al. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLoS Genetics. 2015; 11 (4): e1004969. PubMed PMID: PMC4388571.

Nagamine Y., Pong-Wong R., Navarro P., Vitart V., Hayward C., et al. Localising Loci underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. PLoS ONE. 2012; 7 (10): e46501.

Neill, S. D., Cassidy, J., Hanna, J., Mackie, D. P., Pollock, J. M., et al. Detection of Mycobacterium bovis infection in skin test-negative cattle with an assay for bovine interferon-gamma. Veterinary Record. 1994; 135 (6): 134-135.

O'Neill P. D., Roberts G. O. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1999; 162 (1): 121-129.

Oberley-Deegan R. E., Regan E. A., Kinnula V. L., Crapo J. D. Extracellular Superoxide Dismutase and Risk of COPD. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2009; 6 (4): 307-312. PubMed PMID: 19811392.

Odegard J., Jensen J., Madsen P., Gianola D., Klemetsdal G., et al. Detection of mastitis in dairy cattle by use of mixture models for repeated somatic cell scores: a Bayesian approach via Gibbs sampling. J Dairy Sci. 2003; 86 (11): 3694-3703. Epub 2003/12/16. PubMed PMID: 14672200.

Odegard J., Madsen P., Gianola D., Klemetsdal G., Jensen J., et al. A Bayesian threshold-normal mixture model for analysis of a continuous mastitis-related trait. J Dairy Sci. 2005; 88 (7): 2652-2659. Epub 2005/06/16. PubMed PMID: 15956327.

Odegard J., Meuwissen T. H., Heringstad B., Madsen P. A simple algorithm to estimate genetic variance in an animal threshold model using Bayesian inference. Genet Sel Evol. 2010; 42 (29): 1297-9686.

Olea-Popelka F. J., White P. W., Collins J. D., O'Keeffe J., Kelton D. F., et al. Breakdown severity during a bovine tuberculosis episode as a predictor of future herd breakdowns in Ireland. Prev Vet Med. 2004; 63 (3-4): 163-172.

Ostersen T., Christensen O. F., Henryon M., Nielsen B., Su G., et al. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genetics, Selection, Evolution : GSE. 2011; 43 (1): 38-38. PubMed PMID: PMC3354418.

Pausch H., Aigner B., Emmerling R., Edel C., Gotz K. U., et al. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel Evol. 2013; 45 3. Epub 2013/02/15. PubMed PMID: 23406470; PubMed Central PMCID: PMCPmc3598996.

Phillips C. J. C., Foster C. R. W., Morris P. A., Teverson R. Genetic and management factors that influence the susceptibility of cattle to Mycobacterium bovis infection. Animal Health Research Reviews. 2002; 3 (01): 3-13.

Pollock J. M., Neill S. D. Mycobacterium bovis infection and tuberculosis in cattle. Veterinary journal (London, England : 1997). 2002; 163 (2): 115-127.

Quilez J., Martínez V., Woolliams J. A., Sanchez A., Pong-Wong R., et al. Genetic Control of Canine Leishmaniasis: Genome-Wide Association Study and Genomic Selection Analysis. PLoS One. 2012; 7 (4): e35349.

Qureshi T., Templeton J. W., Adams L. G. Intracellular survival of Brucella abortus, Mycobacterium bovis BCG, Salmonella dublin, and Salmonella typhimurium in macrophages from cattle genetically resistant to Brucella abortus. Vet Immunol Immunopathol. 1996; 50 (1-2): 55-65.

Radunz B. L., Lepper A. W. Suppression of skin reactivity to bovine tuberculin in repeat tests. Australian veterinary journal. 1985; 62 (6): 191-194. Epub 1985/06/01. PubMed PMID: 3904702.

Raman K., Bhat A. G., Chandra N. A systems perspective of host-pathogen interactions: predicting disease outcome in tuberculosis. Mol Biosyst. 2010; 6 (3): 516-530.

Riek R. Studies on the reactions of animals to infestation with ticks. VI. Resistance of cattle to infestation with the tick Boophilus microplus (Canestrini). Australian Journal of Agricultural Research. 1962; 13 (3): 532-550.

Riggio V., Abdel-Aziz M., Matika O., Moreno C. R., Carta A., et al. Accuracy of genomic prediction within and across populations for nematode resistance and body weight traits in sheep. Animal. 2014; 8 (4): 520-528. Epub 2014/03/19. PubMed PMID: 24636823.

Riggio Valentina O. m., Ricardo Pong-Wong, Michael James Stear, and Stephen Christopher Bishop. Genome-Wide Association and Regional Heritability Mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lamps. Submitted to Heredity. 2012

Roughsedge T., Brotherstone S., Visscher P. M. Quantifying genetic contributions to a dairy cattle population using pedigree analysis. Livestock Production Science. 1999; 60 (2): 359-369.

Sales J., Hill W. G. Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. Animal Science. 1976; 23 (01): 1-14.

Sanchez-Molano E., Pong-Wong R., Clements D. N., Blott S. C., Wiener P., et al. Genomic prediction of traits related to canine hip dysplasia. Frontiers in Genetics. 2015; 6.

Sanna S., Jackson A. U., Nagaraja R., Willer C. J., Chen W. M., et al. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet. 2008; 40 (2): 198-203.

Saunders B. M., Cooper A. M. Restraining mycobacteria: role of granulomas in mycobacterial infections. Immunology and cell biology. 2000; 78 (4): 334-341. Epub 2000/08/18. PubMed PMID: 10947857.

Schaeffer L. R. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006; 123 (4): 218-223.

Sequencing T. B. G., Consortium A., Elsik C. G., Tellam R. L., Worley K. C. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science. 2009; 324 (5926): 522-528.

Serão N. V., Matika O., Kemp R. A., Harding J. C., Bishop S. C., et al. Genetic analysis of reproductive traits and antibody response in a PRRS outbreak herd. J Anim Sci. 2014; 92 (7): 2905-2921. Epub 2014/06/01. PubMed PMID: 24879764.

Settles M., Zanella R., McKay S. D., Schnabel R. D., Taylor J. F., et al. A whole genome association analysis identifies loci associated with Mycobacterium avium subsp. paratuberculosis infection status in US holstein cattle. Anim Genet. 2009; 40 (5): 655-662.

Sheridan M. Progress in tuberculosis eradication in Ireland. Vet Microbiol. 2011; 151 (1-2): 160-169.

Simm G. Genetic improvement of cattle and sheep. Ipswich: Farming Press; 1998. p xiii + 433 pp.

Smith E. M., Hoffman J. I., Green L. E., Amos W. Preliminary association of microsatellite heterozygosity with footrot in domestic sheep. Livestock Science. 2012; 143 (2–3): 293-299.

Smith N. H., Berg S., Dale J., Allen A., Rodriguez S., et al. European 1: a globally important clonal complex of Mycobacterium bovis. Infect Genet Evol. 2011; 11 (6): 1340-1351.

Smith N. H., Dale J., Inwald J., Palmer S., Gordon S. V., et al. The population structure of Mycobacterium bovis in Great Britain: clonal expansion. Proc Natl Acad Sci U S A. 2003; 100 (25): 15271-15275.

Smith N. H., Gordon S. V., de la Rua-Domenech R., Clifton-Hadley R. S., Hewinson R. G. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Micro. 2006; 4 (9): 670-681.

Solberg T. R., Sonesson A. K., Woolliams J. A., Meuwissen T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. Genetics, Selection, Evolution : GSE. 2009; 41 (1): 29-29. PubMed PMID: PMC2671482.

Sorokina E. M., Feinstein S. I., Milovanova T. N., Fisher A. B. Identification of the amino acid sequence that targets peroxiredoxin 6 to lysosome-like structures of lung epithelial cells. Am J Physiol Lung Cell Mol Physiol. 2009; 297 (5): 21.

Spencer C. C. A., Su Z., Donnelly P., Marchini J. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLoS Genet. 2009; 5 (5): e1000477.

Stein C. M., Guwatudde D., Nakakeeto M., Peters P., Elston R. C., et al. Heritability Analysis of Cytokines as Intermediate Phenotypes of Tuberculosis. Journal of Infectious Diseases. 2003; 187 (11): 1679-1685.

Stein C. M., Nshuti L., Chiunda A. B., Boom W. H., Elston R. C., et al. Evidence for a Major Gene Influence on Tumor Necrosis Factor-α Expression in Tuberculosis: Path and Segregation Analysis. Human Heredity. 2005; 60 (2): 109-118.

Stein C. M., Zalwango S., Malone L. L., Won S., Mayanja-Kizza H., et al. Genome Scan of M. tuberculosis Infection and Disease in Ugandans. PLoS ONE. 2008; 3 (12): e4094.

Strain S. A. J. M. J., W. McDowell S. J. . Bovine tuberculosis: A review of diagnostic tests for M. bovis infection in cattle. Afbi. 2011.

Su G., Madsen P., Nielsen U. S., Mantysaari E. A., Aamand G. P., et al. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J Dairy Sci. 2012; 95 (2): 909-917. Epub 2012/01/28. PubMed PMID: 22281355.

Suttle N. F., Jones D. G., Woolliams C., Woolliams J. A. Heinz body anaemia in lambs with deficiencies of copper or selenium. The British journal of nutrition. 1987; 58 (3): 539-548. Epub 1987/11/01. PubMed PMID: 3689753.

Team R. D. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 2008. The Danish Pig Research Centre D. A. F. C. Annual Report 2014. 2014.

The State Veterinary Service D. National Scrapie Plan for Great Britain. 2006.

Thom M., Morgan J. H., Hope J. C., Villarreal-Ramos B., Martin M., et al. The effect of repeated tuberculin skin testing of cattle on immune responses and disease following experimental infection with Mycobacterium bovis. Vet Immunol Immunopathol. 2004; 102 (4): 399-412. Epub 2004/11/16. PubMed PMID: 15541793.

Thom M. L., Hope J. C., McAulay M., Villarreal-Ramos B., Coffey T. J., et al. The effect of tuberculin testing on the development of cell-mediated immune responses during Mycobacterium bovis infection. Vet Immunol Immunopathol. 2006; 114 (1-2): 25-36. Epub 2006/08/15. PubMed PMID: 16904754.

Todd D. L., Woolliams J. A., Roughsedge T. Gene flow in a national cross-breeding beef population. animal. 2011; 5 (12): 1874-1886.

Tsairidou S., Woolliams J. A., Allen A. R., Skuce R. A., McBride S. H., et al. Genomic prediction for tuberculosis resistance in dairy cattle. PLoS One. 2014; 9 (5): e96728. Epub 2014/05/09. PubMed PMID: 24809715; PubMed Central PMCID: PMCPmc4014548.

Uemoto Y., Pong-Wong R., Navarro P., Vitart V., Hayward C., et al. The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. Frontiers in Genetics. 2013; 4.

Vanin E. F. Processed pseudogenes: characteristics and evolution. Annual review of genetics. 1985; 19 253-272. Epub 1985/01/01. PubMed PMID: 3909943.

VanRaden P. M. Efficient methods to compute genomic predictions. J Dairy Sci. 2008; 91 (11): 4414-4423.

VanRaden P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., et al. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science. 2009; 92 (1): 16-24.

Vidal S. M., Malo D., Vogan K., Skamene E., Gros P. Natural resistance to infection with intracellular parasites: isolation of a candidate for Bcg. Cell. 1993; 73 (3): 469-485. Epub 1993/05/07. PubMed PMID: 8490962.

Villa-Angulo R., Matukumalli L., Gill C., Choi J., Van Tassell C., et al. High-resolution haplotype block structure in the cattle genome. BMC Genetics. 2009; 10 (1): 19. PubMed PMID: doi:10.1186/1471-2156-10-19.

Villanueva B., Pong-Wong R., Fernandez J., Toro M. A. Benefits from marker-assisted selection under an additive polygenic genetic model. J Anim Sci. 2005; 83 (8): 1747-1752.

Visscher P. M., Medland S. E., Ferreira M. A. R., Morley K. I., Zhu G., et al. Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet. 2006; 2 (3): e41.

Vogels M. W., van Balkom B. W., Heck A. J., de Haan C. A., Rottier P. J., et al. Quantitative proteomic identification of host factors involved in the Salmonella typhimurium infection cycle. Proteomics. 2011; 11 (23): 4477-4491.

Vordermeier M., Gordon S. V., Hewinson R. G. Mycobacterium bovis antigens for the differential diagnosis of vaccinated and infected cattle. Vet Microbiol. 2011; 151 (1-2): 8-13. Epub 2011/03/18. PubMed PMID: 21411245.

Waddington K. To stamp out "so terrible a malady": bovine tuberculosis and tuberculin testing in Britain, 1890-1939. Medical history. 2004; 48 (1): 29-48.

Wang Y., Zhou X., Lin J., Yin F., Xu L., et al. Effects of Mycobacterium bovis on monocyte-derived macrophages from bovine tuberculosis infection and healthy cattle. FEMS Microbiology Letters. 2011; 321 (1): 30-36.

Waters W. R., Palmer M. V., Thacker T. C., Davis W. C., Sreevatsan S., et al. Tuberculosis immunity: opportunities from studies with cattle. Clinical & developmental immunology. 2011; 2011.

Wei W. H., Hemani G., Gyenesei A., Vitart V., Navarro P., et al. Genome-wide analysis of epistasis in body mass index using multiple human populations. European journal of human genetics : EJHG. 2012; 20 (8): 857-862. Epub 2012/02/16. PubMed PMID: 22333899; PubMed Central PMCID: PMCPmc3400731.

Widdison S., Watson M., Piercy J., Howard C., Coffey T. J. Granulocyte chemotactic properties of M. tuberculosis versus M. bovis-infected bovine alveolar macrophages. Molecular Immunology. 2008; 45 (3): 740-749.

Wiener G., Lee G. J., Woolliams J. Effects of rapid inbreeding and of crossing of inbred lines on conception rate, prolificacy and ewe survival in sheep. Animal Production. 1992; 55 115-121.

Wilk J. B., Walter R. E., Laramie J. M., Gottlieb D. J., O'Connor G. T. Framingham Heart Study genome-wide association: results for pulmonary function measures. BMC medical genetics. 2007; 8 Suppl 1 S8. Epub 2007/10/16. PubMed PMID: 17903307; PubMed Central PMCID: PMCPmc1995616.

Willer C. J., Sanna S., Jackson A. U., Scuteri A., Bonnycastle L. L., et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008; 40 (2): 161-169.

Wood Z. A., Schroder E., Robin Harris J., Poole L. B. Structure, mechanism and regulation of peroxiredoxins. Trends Biochem Sci. 2003; 28 (1): 32-40.

Woolliams C., Suttle N. F., Woolliams J., Jones D. G., Wiener G. Studies on lambs from lines genetically selected for low and high copper status. 1. Differences in mortality. Animal Production Science. 1986; 43: 293-301.

Woolliams J., Woolliams C., Suttle N. F., Jones D. G., Wiener G. Studies on lambs from lines genetically selected for low and high copper status. 2. Incidence of hypocuprosis on improved hill pasture. Animal Production Science. 1986; 43 303-317.

Wray N. R., Goddard M. E., Visscher P. M. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007; 17 (10): 1520-1528.

Wray N. R., Yang J., Goddard M. E., Visscher P. M. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. PLoS Genet. 2010; 6 (2): e1000864.

Wray N. R., Yang J., Hayes B. J., Price A. L., Goddard M. E., et al. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013; 14 (7): 507-515.

Yamakawa Y., Pennelegion C., Willcocks S., Stalker A., MacHugh N., et al. Identification and functional characterization of a bovine orthologue to DC-SIGN. Journal of Leukocyte Biology. 2008; 83 (6): 1396-1403.

Yang J., Manolio T. A., Pasquale L. R., Boerwinkle E., Caporaso N., et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43 (6): 519-525. Epub 2011/05/10. PubMed PMID: 21552263; PubMed Central PMCID: PMCPmc4295936.

Zaitlen N., Kraft P., Patterson N., Pasaniuc B., Bhatia G., et al. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet. 2013; 9 (5): e1003520.

Zerehdaran S., van Grevehof E. M., van der Waaij E. H., Bovenhuis H. A bivariate mixture model analysis of body weight and ascites traits in broilers. Poultry science. 2006; 85 (1): 32-38. Epub 2006/02/24. PubMed PMID: 16493943.