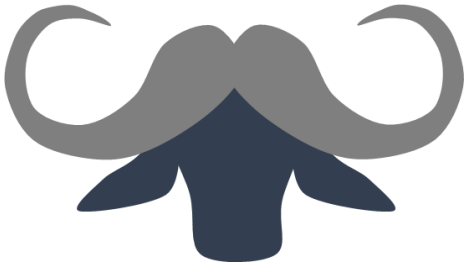


SNP discovery and selection in Cape buffalo for bTB association study, using an African buffalo genome reference

Marius Adriaan Engelbrecht

Thesis presented in partial fulfilment of the requirements for the degree Master of Science in Human Genetics in the Faculty of Medicine and Health Sciences at the University of Stellenbosch

Supervisor: Prof. Craig Kinnear

Co-supervisors: Prof. Marlo Moller; Dr. Brigitte Glanzmann

Faculty of Medicine and Health Sciences, Department Biomedical Sciences, Division of Molecular Biology and Human Genetics

March 2021

Declaration

By submitting this thesis/dissertation, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third-party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature ……..

Date: ……..

Acknowledgements

Firstly, and most importantly, I would like to acknowledge my Supervisors, Prof. Craig Kinnear, Prof. Marlo Moller and Dr. Brigitte Glanzmann for all they have done for me and this thesis. I whole heartedly appreciate all of the time and effort they have put in to making this project possible. I am very grateful for the opportunity to work with such dedicated individuals. I would like to express my gratitude for all the help throughout the course of this project, and for being patient with me regardless of the struggles that were faced. The guidance, advice and technical aid provided was integral to this endeavour.

I would also like to thank Dr. Brigitte Glanzmann for providing the reference genome for the Cape Buffalo which she produced. This formed a basis for this project and thus was an indispensable part of it.

I would also like to acknowledge Dr. Deon De Jager for allow me to access and make use of his previous work and data, as well as help with the project. Without this input this project would not have been possible, and I would not have had such an opportunity to do this work.

Further I would like to express my appreciation to everyone in the department and research group that offered assistance and helped me with the project in any way they could. Specifically, Prof. David Tabb, Ms. Tina Meiring and Ms. Anel Sparks for all the help with the bioinformatic and computational work.

I especially would like to extend my gratitude to my good friend and colleague, Robin G Swart. We faced all our challenges together and helped each with almost everything we could. These years would not have been the same without him, and I am glad we made it through this together.

I would like to acknowledge Stellenbosch University, The Centre of Excellence for Biomedical Tuberculosis Research (CBTBR), and the National Research Foundation (NRF) for financially supporting this project and me as an individual for the duration of said project.

Lastly, I would like to acknowledge my family for always supporting me, no matter what happened.

Abbreviations

AENP	Addo Elephant National Park
BAM	Binary Alignment/Map (file)
BCG	Bacille Calmette Guerin
BLASTn	Basic Local Alignment Search Tool nucleotide
bTB	Bovine tuberculosis
BWA	Burrows-Wheeler Aligner
CG	Combined Genotyping
CGDR	Combined Genotyping with duplicates removed
CI	confidence intervals
CMI	cell-mediated immune (response)
EKZNW	Ezemvelo KwaZulu Natal Wildlife
GAS	Genetic Association Study
GATK	Genome Analysis Toolkit
Gb	gigabase
GWAS	genome wide association study
HiP	Hluhluwe-iMfolozi Park
HWE	Hardy-Weinberg equilibrium
IFN-γ	interferon gamma
IG	Individual Genotyping
IGRA	interferon gamma release assay
IL-1a	interleukin-1-alpha
INDEL(s)	insertion(s)/deletion(s)
IP-10	Interferon gamma-inducible protein 10
IUCN	Union for Conservation of Nature
kb	kilobases
KNP	Kruger National Park
LD	linkage disequilibrium
MAC	minor allele count
MAF(s)	minor allele frequencie(s)
Mb	megabases
MHC	major histocompatibility complex
MNP	Mokala National Park
NGS	next generation sequencing
NK	natural killer cell(s)
OR	odds-ratio
PCA	principal component analysis
PCR	polymerase chain reaction
PPD	Purified Protein Derivative
QFT	QuantiFERON ®-TB Gold

Q-Q	Quantile-Quantile
RT-qPCR	Real-Time quantitative polymerase chain reaction
SAM	Sequence Alignment/Map (file)
SCITT	single comparative intradermal tuberculin test
SLC11A1	solute carrier family 11 member 1
SNP(s)	single nucleotide polymorphism(s)
TB	tuberculosis
TST	tuberculin skin test
UK	United Kingdom
USA	United States of America
VCF	Variant Call Format (file)
VEP	Variant Effect Predictor
WGS	whole genome sequencing
WHO	World Health Organization

ABSTRACT

The African buffalo (*Syncerus caffer*) is an important herd-based bovid in Africa, which is ubiquitous across almost the entire continent. These animals also act as a maintenance host for the ever-present threat that is bovine tuberculosis (bTB). The animal facilitates the spread and continued existence of the health problem that is bTB amongst wildlife and domestic cattle populations throughout Africa, causing problems in terms of conservation and economic loss. The disease is endemic to the southern part of Africa, especially South Africa, where two major national parks, The Kruger National Park (KNP) and Hluhluwe-iMfolozi Park (HiP), are host to it. There are also spill-over events of the disease from animals to humans, which is especially problematic in South Africa where tuberculosis (TB) in humans is already a major health concern. This study aimed to use 40 high-quality low-coverage African buffalo whole genome sequences in conjunction with a species-specific reference genome to create a panel of single nucleotide polymorphisms (SNPs) for use in further research in genetic association in buffalo bTB susceptibility. The sequences were from 40 Cape buffalo from 4 South African national parks, namely KNP, HiP and two bTB unexposed regions, the Mokala National Park (MNP) and Addo Elephant National Park (AENP). From this we produced a panel of 3698 high quality SNPs across 26 immune related genes in the African buffalo genome. One hundred and forty-three of these SNPs in three genes from the panel was used in a preliminary targeted association test with bTB exposure, which produced 10 SNPs associated with TB exposure. This may aid in future research and subsequent association studies.

CHAPTER 1:

INTRODUCTION

Contents

1.1 Summary

The African buffalo (*Syncerus caffer; S. c. spp*) is not just a member of the “Big Five”, it is an animal which affects and influences other animals and humans in several important ways. For one, African buffalo are susceptible to bovine tuberculosis (bTB) and is able to spread this disease to both other wild animals and domestic livestock [1,2]. Additionally, the African buffalo spread tuberculosis (TB) to humans [1]. Apart from the obvious health risk this causes, bTB has a negative effect on the tourism and agricultural sectors in South Africa, causing severe economic consequences [3–6]. To formulate effective methods of bTB control, increased focus has been placed on investigating the host genetic factors that contribute to the development of bTB. With the advancement and reduction cost for genetic sequencing technologies it has become more viable to do high resolution studies on a genome wide scale, especially in non-model organisms. This can be seen in the genetic research that have been done on African buffalo in the last two decades.

1.2 African buffalo species

African buffaloes are found in large portions of sub-Saharan Africa (Figure 1.1) and are divided into three main subspecies. The Cape buffalo (*Syncerus caffer caffer*) is the predominant sub species of African buffalo in southern Africa and is the species which is associated with the “Big Five” game animals in Africa. The Forest buffalo (*Syncerus caffer nanus*) and the West African buffalo (*Syncerus caffer brachyeros*) both inhabit west and central Africa, where *S. c. nanus* is generally found in the more southern of the two regions. The African buffalo is considered to be the largest bovid species to be found in the African savanna ecosystem with an average weight of approximately 450kg [7,8].

African buffalo are herd based bulk feeding grazers and are of great ecological importance to this ecosystem. Through their feeding pattern, these animals influence that of other grazers in the savanna [8].

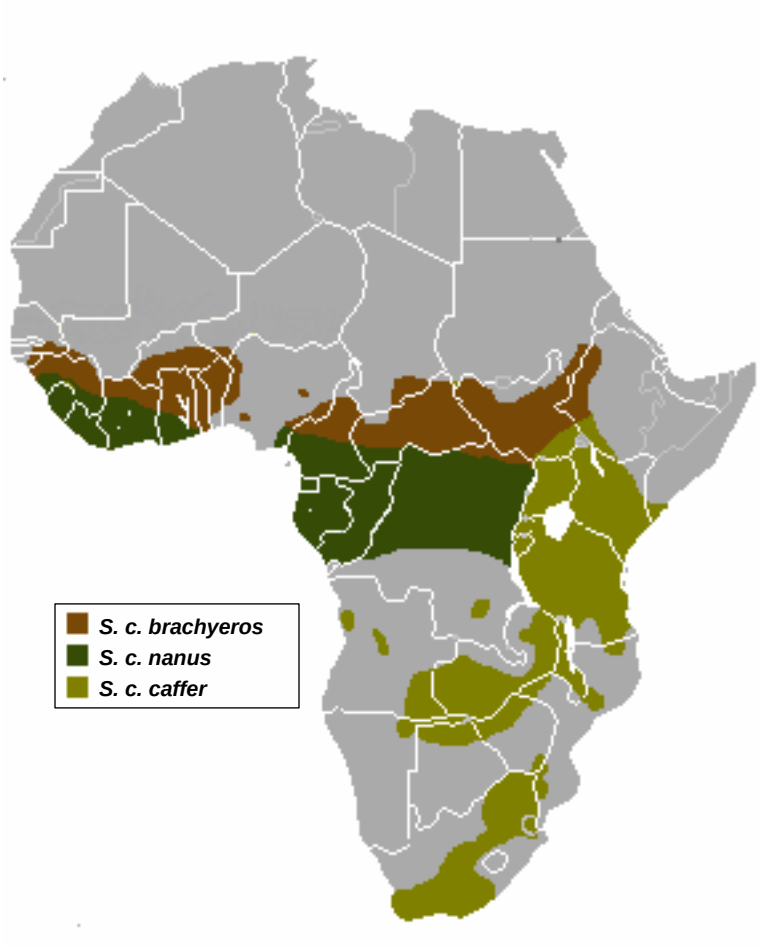


Figure 1.1: The distribution of *Syncerus caffer* and its sub-species in Africa. Image adapted from IUCN SSC Antelope Specialist Group 2019 (http://dx.doi.org/10.2305/IUCN.UK.2019-1.RLTS.T21251A50195031.en)

Considering the importance of this species, it is concerning that the population of African buffalo has been on the decline, with the majority (~75%) of Cape buffalo located in protected wildlife areas in 2014 [9]. According to the International Union for Conservation of Nature (IUCN) census in 2011, there were approximately 900 000 African buffalo

globally, while the most recent IUCN numbers from 2018 report around 398 000 to 401 000 mature individuals remaining, with the IUCN status of being 'near threatened' [10,11]. Aside from humans being a threat to these animals through poaching and expansion, African buffalo also have to contend with bTB.

1.3 Bovine Tuberculosis

1.3.1 Background

Bovine tuberculosis (bTB) is a chronic infectious disease and is closely related to the human form of tuberculosis (TB). bTB is primarily caused by the pathogenic bacteria *Mycobacterium bovis*, which is part of the *Mycobacterium tuberculosis* complex (MTBC) along with the predominant causative agent of TB in humans: *Mycobacterium tuberculosis* (*M tuberculosis*). bTB is historically associated with *M. bovis* infections in cattle, however the bacterium has a wide range of mammalian hosts, causing bTB through infection in wild to domesticated animals [12]. Domestic animals include house pets, such as domesticated cats and dogs. Besides bovids, wild animal hosts range from warthogs, wild boar, dear to lions[13–17]. *M. bovis* can also infect humans, and causes what is known as zoonotic TB [18]. A zoonotic disease is a disease which affects animals which can also be transmitted to humans. For the purpose of this thesis, TB in humans caused by *M. bovis* will simply be referred to as bTB. Thus, in the context of the bTB in humans it is implied that infection with *M. bovis* causes zoonotic TB.

M. bovis is acid-fast bacterium and weakly appears Gram-positive with Gram-staining due to a thick outer cell wall which contains Mycolic acid which is characteristic of *Mycobacterium* species [12]. *M. bovis* grows aerobically as a facultative intracellular parasite and, as with *M. tuberculosis*, it also has an extremely slow generation time (16-

20 hours) [12]. These factors coupled with growth requirements such as specialized medium, make it difficult to do culture-based diagnostics and research on these bacteria [12].

1.3.3 Pathogenesis

The disease progression of bTB is summarised in Figure 1.2. Infection with *M. bovis* in cattle and buffalo has a similar disease progression to that of *M. tuberculosis* in humans, in that the bacteria can infect an individual, but remain dormant, not causing any

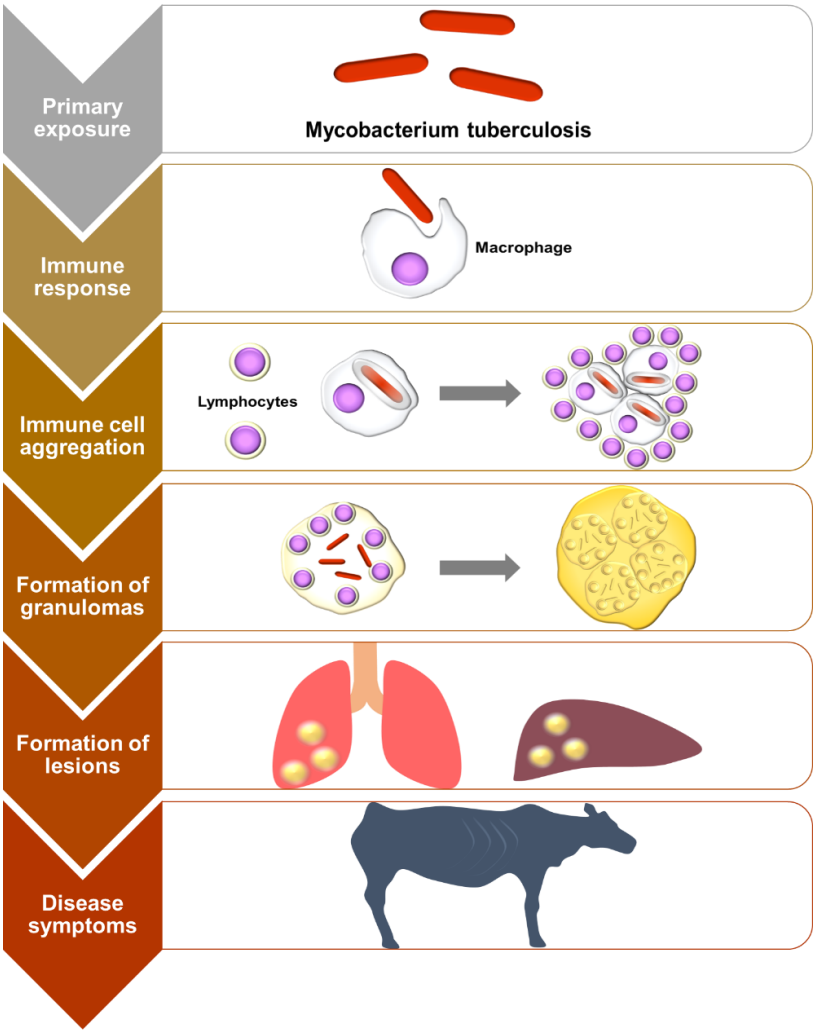


Figure 1.2: Basic representation of the disease progression of tuberculosis (TB) and the life cycle of *Mycobacterium spp*. bacilli in the form of the disease progression of pulmonary TB.

detectable symptoms for extended periods of time [19,20]. The localization of the disease in the body depends on the method of infection/transmission [19]. The host's immune system plays an important role in the progression of the disease, as the initial phases of infection are driven by the primary immune response and subsequent containment of the bacteria. *M. bovis* infection causes the formation of lesions in the infected organs, usually the lung or upper respiratory tract, and progresses to the infection of lymph nodes [19,21].

Figure 1.2 shows the progression of pulmonary TB infection. During the initial infection the host immune system detects the *M. bovis* bacteria and this triggers an inflammatory response followed by the phagocytosis of the bacteria by the macrophages [19,22]. This induces the accumulation of other host immune cells, such as lymphocytes including T cells and natural killer (NK) cells, at the site of infection, resulting in the formation of granulomas: a mass of host cells within the infected tissue [19,21]. These granulomas along with aggregation of immune cells causes the formation of lesions within the infected tissue[19]. These can then mature and spread throughout the infected tissue causing the formation of firm dry caseous necrotic masses[22]. This is known as caseation which is characteristic of TB. The symptoms are similar to TB in humans, such as fever and respiratory complications in the case of pulmonary infection [19]. Additionally, symptoms include weight loss and fatigue, giving the appearance of malnutrition. Various organ function can be impaired based on the localization of the infection and lesions [19].

1.3.4 Transmission

Bovine TB is transmitted in several ways, including genital, ingestion and aerosol [19]. The latter is the most common cause of bTB infection in cattle and buffalo [19,23,24]. Another common method of infection is the consumption of infected feed [14,19,23–25]. In general, the

spread of the disease is associated with close contact with infected individuals. This applies to animals and humans, although in recent history cases of bTB in humans have largely been attributed to the consumption of infected animal products such as meats and unpasteurized milk [26]. The disease can similarly be transmitted to domestic and companion animals through close contact, in the context of an environment where they are in close proximity to a host population, and act as what are known as spill-over hosts [23]. These animals can act as transmission intermediaries between wild animals or cattle and humans [27].

1.4 Prevalence of bTB

1.4.1 Prevalence of bTB in cattle

Bovine TB incidence has decreased since the mid-20th century, due to stringent bTB control campaigns in livestock in the developed world [24,28,29]. However, bTB is still considered a significant health problem with countries such as the United States of America (USA), the United Kingdom (UK) and New Zealand still reporting cases [14,28,30,31]. With the eradication/control of bTB in cattle in much of the developed world, bTB has the greatest incidence in developing countries. Many of these countries lack the finances and infrastructure to facilitate efficient bTB eradication and control programs. Consequently, surveillance and data collection with regards to bTB incidence and other relevant information are inaccurate or in many cases completely absent [16,26,32]. This means that accurate prevalence statics are rare for these regions. The Indian sub-continent is one such region where bTB is present. India makes use of domesticated and undomesticated livestock for agricultural production, hosting the largest population of cattle in the world [33]. The country lacks a formal bTB control program and has an estimated 7.3%

prevalence of bTB in cattle. This equates to a relatively large amount of affected animals, and thus a potentially serious problem if not approached correctly [33].

In Africa, where bTB control programs have yet to be enforced on the level of developed counties, the prevalence of the disease is relatively high. In Ethiopia, a country with approximately 50 million cattle[34], the prevalence of bTB in cattle-herds (depending of the husbandry system) range from 3.4 to 50% [23,34]. Southern Africa has a well-known history of TB, however the prevalence of bTB in cattle is currently poorly documented [16,35]. The presence of several bTB maintenance hosts among the wildlife has made it so that there is the persistent threat of bTB being transmitted to cattle and thus increasing the disease prevalence in the region [16].

Maintaining accurate account of infected versus disease individuals is difficultsince *M. bovis* infection does not necessarily lead to active disease [19,36]. This confounds the collection of accurate statistics for of the amount of communicable individuals. This is especially true for the under-developed, rural regions where the disease is still a major concern. Bovine TB is not well understood in wildlife, thus the effects of the misconception of true prevalence of infected versus diseases individuals is detrimental when trying to control the disease [36].

1.4.2 Prevalence of bTB in humans

Tuberculosis in humans still remains a significant problem in counties where the disease is endemic. In countries such as South Africa where there is a high HIV/TB disease burden the impact of TB caused by *M.bovis* can easily be misinterpreted and underestimated. In 2018 the World Health Organization (WHO) reported that Africa accounts for approximately 24% of new TB cases in the world [33]. In that same year, the

WHO also estimated that Africa accounts for approximately 50% of the world's bTB cases in humans resulting in an estimated 9 270 of the global 12 500 human bTB deaths [33].

1.5 bTB in wildlife

1.5.1 Wildlife reservoirs of bTB

The terms 'wildlife reservoir' and 'maintenance host' in the context of diseases are used interchangeably in literature [37,38]. In general the definition of a maintenance host is a host species in which a pathogen persists indefinitely within a given population [37,38]. The term 'wildlife reservoir' is generally defined similarly to the previous term, but with the added context of specifically occurring in a species of wildlife. The formation of a wildlife reservoir is dependent on several factors, such as the population size and density, as well as the level of intraspecies transmission of the pathogen [38]. Conversely, spill-over hosts manifest when the prevalence of a pathogen in one host species is such that the pathogen is transmitted ('spills over') to another host species [38].

The impact that wildlife has on the epidemiology of bTB has become a more relevant topic of research as the prevalence of the disease decrease in cattle. Initially, during the period when bTB eradication programs were first being implemented, cattle were considered to be the source population [38]. Wildlife species that were infected with the disease were considered to be the target of 'spill-over' infections. This is now no longer the case, and in many circumstances the roles have been reversed. The potential for spill-over of bTB is due to the wide host range of *M. bovis,* thus research on wildlife reservoirs of bTB has become important due to the impact they have on controlling the disease in several instances [39–41].

There are a few species that are accepted to act as wildlife reservoirs of bTB across multiple countries. One such species is the European/Eurasian badger (*Meles meles*) [38]. Several studies have been conducted on the role of *M. meles* in the spread and prevalence of bTB in the UK and Ireland [31,42]. The animal is reported to be a significant contributor to the persistence of bTB in cattle in both countries which lead to the culling of badgers as part of bTB eradication programs [42]. In the USA, the white-tailed deer (*Odocoileus virginianus*) has been found to cause bTB infections in domestic cattle as a result of spill-back [14,43]. Control of the disease has progressed to the point of attempting to use forms of the TB vaccine used in humans, namely the Bacille Calmette Guerin (BCG) vaccine [43]. The use of such vaccines have shown to decrease the severity of the disease in animals, specifically lesion formation, yet it does not provide guarantied immunity to the disease [43]. New Zealand hosts the Brushtail possums (*Trichosurus vulpecula*) which is also a target of bTB control efforts due to its role as a wildlife reservoir of the disease [41]. These countries invest considerable resources towards the study of these animals relative to the more disease prevalent regions such as Africa.

1.5.2 bTB in African buffalo and African wildlife

Africa is home to multiple species of wild animals that are affected by bTB, one of which is considered to be a primary bTB maintenance host, namely the African buffalo [16,44]. *S. caffer* is a ubiquitous species in Africa with a wide range of habitats, giving it the potential for spreading bTB to a wide array of spill-over hosts and potentially humans. The Kruger National Park (KNP) and Hluhluwe-iMfolozi Park (HiP) in South Africa are bTB endemic regions. This means these national wildlife parks are regions where animals with bTB are prevalent within their populations, as well as containing maintenance hosts of the disease.

The disease was introduced into the parks via contact between the wildlife and cattle from local communal farms [32]. Since its introduction, bTB has spread to several wildlife species in these regions. For example, apart from *S. caffer*, the disease has been found in, lions (*Panthera leo*), white rhinoceros (*Ceratotherium simum*) and black rhinoceros (*Diceros bicornis*) [17,45,46]. A recent study conducted in South Africa found evidence suggesting that bTB has spilled over to16 different animals species via inter species transmission of *M. bovis* [47], which has increased to 24 different animals [36]. Interestingly, *M.bovis* and *M.tuberculosis* have also been isolated form African elephant (*Loxodonta africana*) individuals [48]. This naturally has significant implications for the ecology and conservation of effected species. Restricting the spread of bTB in animals such as the African buffalo is clearly of great importance to control the disease.

1.6 Economic impact of bTB

1.6.1 Global economic impact

Bovine TB can have a significant negative impact on not only an ecological level but an economic one as well. There are several expenses tied to the surveillance and control of the disease. Most developed countries invest heavily in bTB control programs. Since 2005, the USA annual federal spending on their bTB control program was approximately $15 million annually [18]. An additional $200 million was spent as emergency funding on the program for research, development, eradication and control between 2000 and 2008 [18]. Ireland and the UK spends an estimated $70 million and $1.5 billion per year on their bTB control program respectively, as of May 2015 [18]. Globally the disease is estimated to incur $3 billion in agricultural losses annually [3].

1.6.2 Economic impact bTB in Africa

Since it threatens the animals directly tied to the very image of African wildlife there is no doubt that improper control of bTB can have a devastating effect on eco-tourism in countries like South Africa. Privately owned game farms and trophy hunting are also large contributors to the South African economy [27], where the trade of live game was predicted to contribute approximately $2 billion each year [6]. On auction, trophy quality Cape buffalo are in high demand for breeding purposes. African buffalo alone accounted for an estimated ~$0.7 billion to the South African trophy hunting industry in 2015[6] The presence of bTB in commercial livestock has a devastating economic effect, since it results in the cost of identifying the extent of the disease in the herds and the loss of productivity and animals due to the culling of the diseased individuals and quarantining affected herds [49].

1.7 bTB control and diagnosis

Effective control of diseases such as bTB relies heavily on the diagnosis, and accurate tracking of infected populations. Initially detecting the disease in infected animals is difficult because it is a chronic infection where clinical signs of the disease can be absent for a prolonged period of time [19,50]. Most bTB control programs in commercial cattle are based on routine screening and testing of animals [3,49,51]. This is usually performed using methods that measure antigen-specific cell-mediated immune (CMI) responses to detect infection. One method typically used in cattle is the tuberculin skin test (TST). The test measures the CMI response of an individual after the intradermal injection of tuberculin. The response is measured between 48 and 72 hours after the injection. CMI manifests itself as an induration of the skin which is measured to determine the TB status of the individual. The methods and practices of physically measuring the size of these

indurations lends itself to operator bias for the test results. Tuberculin is a Mycobacterial Purified Protein Derivative (PPD) from the specific species of bacteria being tested for. TST is considered to be flawed in some ways, one being that it has a false negative rate of approximately 20% [50]. Another diagnostic test using an *in vitro* approach is the interferon gamma (IFN-γ) release assay (IGRA). This test was the subsequent test to be developed and used. It measures the production of IFN-γ in whole blood samples taken from the individual, which is stimulated with tuberculin. IFN-γ is a cytokine which plays an important role in TB immunity as it forms part of several immune pathways [21,32,50]. IFN-γ can increase the lysosomal activity of macrophages in granulomas which aids in the clearance of intracellular pathogens [21,32,50]. The cytokine is primarily secreted by CD4+ T cells (T helper cells), and in the case of TB infection, these cells will reactively produce high levels of IFN-γ in the presence of mycobacterial components [22,44,52]. Thus trough the stigmatisation whole blood with tuberculin, T cells from TB infected individuals should produce high levels of IFN-γ, and thus can be observed during a IGRA tests, and as such a diagnosis can be made [22,44,52]. The IGRA tests are less likely to be influenced by operator bias compared to TST as it is based on a standardised system of measurement. There is however an increased chance of false positives as a result of cross reaction with non-TB mycobacteria present in the host [52]. The Bovigam® PPD assay is one such test, which uses an enzyme-linked immunosorbent assay (ELISA) to detect the presence of INF-γ [36]. One of the latest wide-spread diagnostic test to be developed from this method and used for bTB is the QuantiFERON ®-TB Gold (QFT) platform which tests both INF-γ and Interferon gamma-inducible protein 10 (IP-10), with ELISAs specifically designed and validated for the host species [36].

The gold standard for TB diagnosis is a culture test, but this also has its limitations, primarily the long time it takes to produce results (6 weeks) [44].

One of the challenges regarding the use of tuberculin, is that diagnostic tests are specific to the species that is being tested [22,44,52]. This means there is limited scope for the application of these types of diagnostics in terms of affected species. Containing the disease in species which can be more easily diagnosed, such as African buffalo, is important. As of the time of writing, there is only one ante-mortem diagnostic test for bTB in African Buffalo approved in South Africa [52]. The test a single comparative intradermal tuberculin test (SCITT), which is a type of TST. It uses PPDs derived from both *M. bovis* and *Mycobacterium avium* [52]. New diagnostic solutions for TB and bTB have been a focus for research and development for decades, and diagnosis of bTB in African buffalo is no exception. Recently two alternate methods have been tested in buffalo in South Africa. Bernitz et al [52] used the QuantiFERON ®-TB Gold (QFT) and showed that the method improved and simplified the detection of bTB in African buffalo. To circumvent the species specificity of CMI based methods, the method tested by Goosen et al[44] took a different approach. Their method used Real-Time quantitative polymerase chain reaction (RT-qPCR) to detect Mycobacterial DNA in animal blood samples. The use of DNA in disease diagnostics is becoming more established with the advent of next generation sequencing (NGS). As these sequencing methods have become more affordable and accessible, so too does the possibility increase for improving our understanding diseases such as bTB and their hosts.

1.8 bTB host genetics

Host genetics is a crucial component of bTB susceptibility [30,53–56]. However, due to the complex nature of TB, and the fact that the pathogens that cause it are yet to be fully understood, there is much that is still unknown. While several investigations have been conducted to investigate the role of host genetics in human TB susceptibility, there has been a dearth of studies investigating genetic susceptibility to TB in animals. There have, however been some work investigating heritability of genetic traits that contribute to bTB resistance in cattle [51,54,57]. The rationale behinds this is that if the traits that increase resistance can be inherited, then these traits can be selected for, thereby producing bTB resistant populations. This was used as the basis to develop an alternative method of disease control in countries such as the UK and Ireland, where bTB remains a notable problem [18,54,58].

With the emergence of more sequencing technologies, the use of gene-based association tests has become more common, phasing out heritability analysis-based techniques. However, conducting association testing in animals is hampered by the same issues as conducting these analyses in humans. There are multiple factors that confound the identification of true associations, many of which can be attributed to the complexity of TB disease [56]. This method relies on the identification of host genetic loci that either increases or decreases the susceptibility to disease [51,59]. There are two main approaches to finding associated genes. The first is a targeted approach, whereby a panel of candidate genes are selected and subsequently used for an association analysis. The second uses whole genome/exome and genotyping arrays sequences from individuals to perform what is known as a genome wide association study (GWAS).

There are multiple genes, in humans (Figure 1.3) and animals that have been associated with TB infection and susceptibility. These genes are commonly associated with immunity

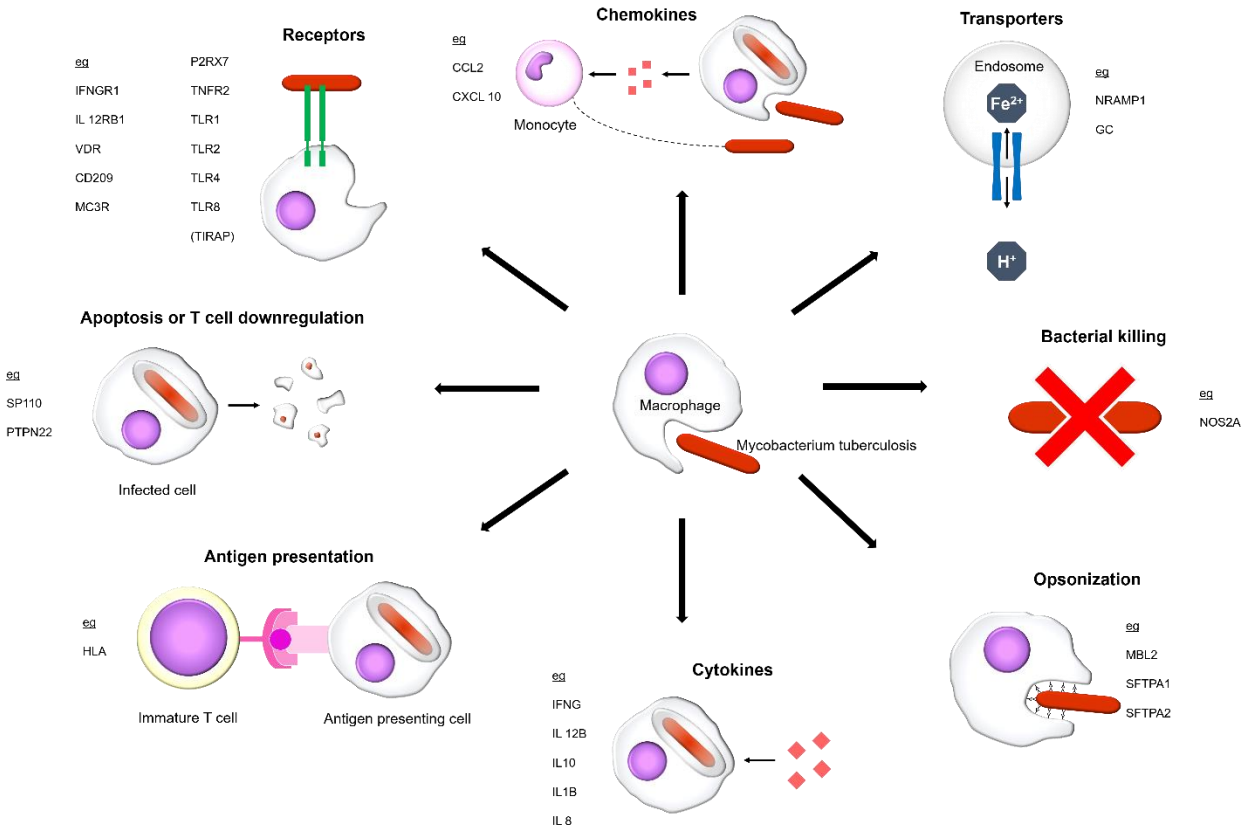


Figure 1.3: Examples of the genes and associated pathways that have recently been found to be associated with TB infection in humans. Adapted from Möller et al. 2010.

or the function of immune pathways, directly or indirectly [56,60]. Functions such as pathogen detection, cytokine production/reception and membrane transport are readily under scrutiny when it comes the molecular analysis. Variations in the genes that contribute to these functions are by extension common targets for research [56,61–63].

Many of the genes that have been implicated in disease susceptibility in animals are orthologous (serves the same function as) genes found in humans with similar associations. Recently the immune genes: *BoLA-DRB3* (cow equivalent to major histocompatibility complex (MHC) gene in humans) and *SLC11A1* have been investigated

and associated, respectively with bTB susceptibility in cattle [64,65]. Both these genes are involved in major immune pathways, which both have been implicated in TB immunity [64,65]. The wide host range of the pathogens in the MTBC can be indicative of genetic traits that are shared among host species [26,51,63]. This overlap means that the common trends and mechanism in immunity regarding TB infection can be investigated across multiple species which can aid in our understanding the pathology of both TB and bTB.

Genetic association studies for bTB susceptibility have been performed in cattle and other bovids used as livestock. Similar studies have been performed on species that act as maintenance hosts for the disease. However, there has yet to be definitive results that elucidate the extent that host genetics influence bTB susceptibility. There have been several genetic studies of the African buffalo in the past decade, and only a few have been focused on bTB [9,66,67]. There can be multiple limitations to working with a non-model organism such as *S. caffer*, especially on a molecular level. This includes variant data such as annotations, populations frequencies possible gene structures and inheritance information that is absent [7,11,68]. This lack of information means that there are several factors and intricacies that are unknown and thus can confound research. Usually the most similar model organism will be used as a reference point or the basis of any research done on non-model organisms, which can lead to biased results [68–71].

1.9 Previous genetic research on bTB in African buffalo

In a study performed by le Roux et al [11], low-coverage whole genome sequencing (WGS) was performed on DNA from 9 Cape buffalo. The resulting sequences were aligned to the domestic cow (*Bos taurus*) genome as a reference, as there was no *S. caffer*

reference genome available at the time. The cow represented the closet model organism to the African buffalo. The alignment was used to produce a set of selection validated novel single nucleotide polymorphisms (SNPs). These SNPs were then used in a subsequent study by le Roux et al [63] to identify polymorphisms associated with susceptibility to bTB infection in Cape buffalo from the KNP and HiP. After validation, the study found 3 significantly associated SNPs in the African buffalo, all of which were in genes previously associated with TB infection. These three SNPs were located in *SLC7A13*, *DMBT1* and *IL1a* cow genes. In *IL1a,* the SNP fell within the 3' untranslated region of the gene (the SNP had no accession number). While the other two SNPs (Ensembl ID: ENSBTAG00000040461 and ENSBTAG00000022715) were missense mutations within coding regions of the genes [63].

As stated in the latter of the two mentioned studies, the use of a model organism as a reference as opposed to a true species specific reference can have confounding effects on the results [63]. One of the problems encountered in the first study is that the buffalo sequences had very poor mapping rates (~20%) to the cow reference genome, and the mapped sequencing reads coverage was 2.7 x, which can lead to an increase in false positive variants due to the lack of appropriate depth per variant. Even though the results of the study are very compelling, it begs the question of "What would the results look like if a more precise reference genome were used?" The use of an inappropriate reference genome can lead to very poor sequencing alignment and thus a great deal of valuable genetic information can be lost. Species and reference specific traits require high quality alignments to identify, this include genetic variants that can only be found when using the correct reference genome. Finding false positives and negatives are naturally also always

a risk, as certain variants could be wildtype in one reference while being a mutation in the appropriate reference.

The use of the appropriate techniques and genetic data is crucial to understanding the basis of the pathology for a disease such as bTB and TB. Any discovery made in any of the wide range of bTB hosts can further the understanding of the disease in several species, due to the pathological overlap that is present in MTBC pathogens. This includes discoveries relating to the genetic predisposition certain individuals show to developing the disease, since gene ontology can shed light on the host mechanism involved in more than one species. For example, identifying genes and variants associated with immune pathways involved in bTB infection in one species, can lay the foundation for targeted studies across multiple host species. This information can be used to develop host targeted treatments for the disease, and as such it can lead to advancements in the control of TB in not only animals, but humans as well. Ultimately this can help in alleviating the health and economic burden TB places on the countries, such as South Africa and also the world as a whole.

Thesis scope

This thesis presents research pertaining to the identification and selection of single nucleotide polymorphisms (SNPs) that can be used for a targeted genetic association test for susceptibility to bovine Tuberculosis (bTB) infection in African buffalo. It also describes further information and research that should serve as the groundwork for said association test. This was done using whole genome sequences from 40 Cape buffalo, and an African buffalo reference genome produced by Glanzmann et al. in 2016. The sequencing data

was processed to produce population statistics and to identify SNPs using multiple bioinformatics tools.

The study populations used in in this study were from 4 different wildlife parks in South Africa (see Chapter 2). Some of the individuals were included in the previous association studies by Le Roux et al. The cohort was used as a sample of the population that further association tests can be performed on. The use of WGS data from these samples facilitates the identification of genetic variants throughout the entire genome. This aids in the extraction of valuable population statistics, and the discovery of variants of interest in more than just the coding regions of the genome. This is especially useful in a non-model organism like *S. caffer*, as there are few sources of relevant genetic information documented. Variant data was extracted from the raw sequences using multiple software tools and packages. Multiple iterations of processing pipelines were implemented using different configurations and parameters to optimize output speed and quality.

Subsequent data analysis were conducted on the variant data produced. This includes basic population statistics and SNP selection. Statistics such as inbreeding and relatedness were produced, and a principal component analysis was performed to measure population structure. SNPs in genes previously associated with bTB infection in African buffalo and cattle were selected.

The results are described in Chapter 3 and concluding remarks and discussion of the study as a whole are covered in Chapter 4. Topics such as the limitations of the study and the final SNP panels produced, and potential applications and improvements that can be made for future or repeat studies are also discussed.

CHAPTER 2:

MATERIALS AND METHODS

Contents

2.1 Summary

Figure 2.1 provides an overview of the methodology used in the present study.

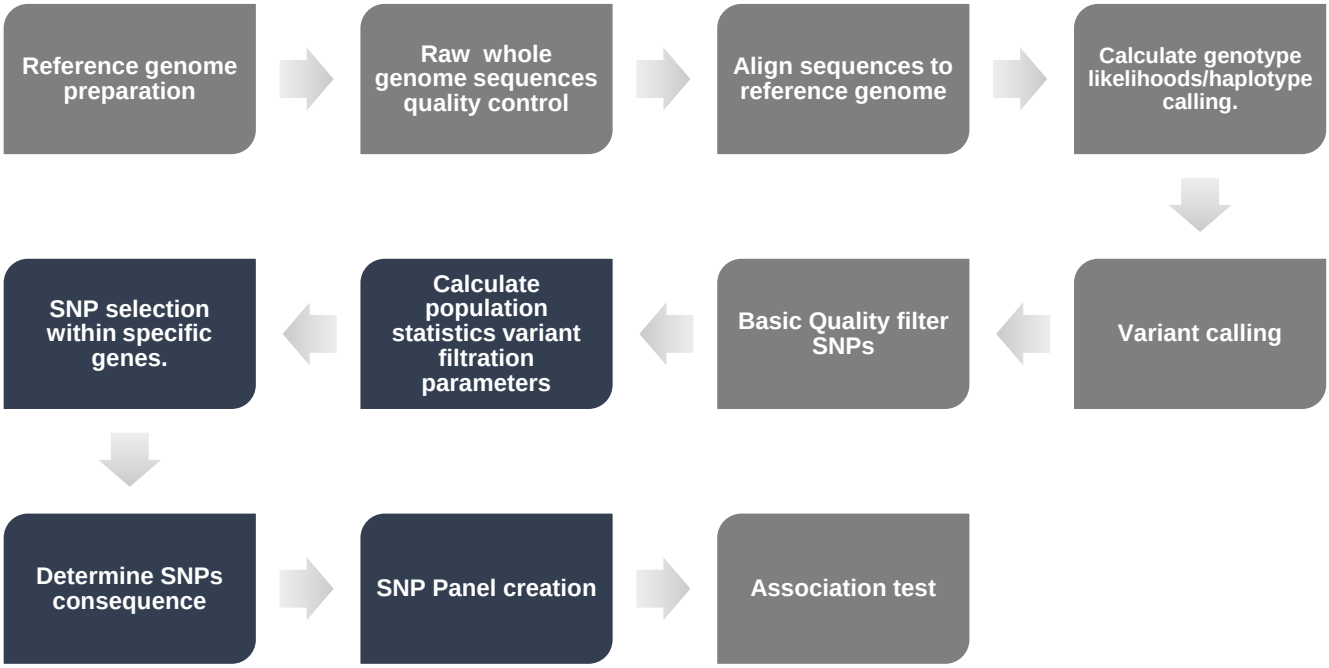


Figure 2.1: A flow diagram depicting the basic structure of the methodology used in this study.

2.2 Samples and DNA extraction

A total of 40 buffalo blood samples were obtained from South African National Parks and Ezemvelo KwaZulu Natal Wildlife (EKZNW). These samples originated from 4 parks: Addo Elephant National Park (AENP) (n=5), Mokala National Park (MNP) (n=5), Kruger National Park (KNP) (n=15) and Hluhluwe-iMfolozi National Park (HiP) (n=15) respectively. At the time of sampling, the individuals sampled from KNP and HiP populations were considered to be bTB exposed, and the samples from the other two parks were bTB unexposed.

High molecular weight DNA was extracted from samples obtained from AENP and MNP using the MagAttract® High Molecular Weight DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. All remaining samples were extracted using the Illustra Nucleon BACC3 RPN8512 Nucleon Kit (GE Healthcare Life Sciences, Chicago, Il, United States), according to the manufacturer's instructions.

2.3 Genome sequencing

Low coverage whole genome sequencing was performed on the 40 buffalo samples by Novagene (Beijing, People's Republic of China). DNA sample preparation was performed using 1.0 µg DNA per sample. The Truseq Nano DNA HT Sample Preparation Kit (Illumina, San Diego, CA, USA) was used to prepare and generate sequencing libraries following the manufacturer's instructions. Sequences were assigned index codes to attribute them to their respective samples. A Covaris sonicator (Covaris, Woburn, MA, USA) was used to randomly fragment the genomic DNA. The resulting 350 base pair (bp) fragments were subsequently processed for Illumina sequencing with subsequent polymerase chain reaction (PCR) amplification. The processing included end polishing, the addition of poly A-tails and ligation with full-length adapters. The AMPure XP system (Beckman Coulter, Brea, CA, USA) was used to purify the PCR products. The Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) and real-time PCR was used to analyze and quantify the size distribution of the libraries, respectively. A HiSeq X instrument (Illumina) was used to perform paired-end sequencing of 150bp. Of the 40 samples, 17 were sequenced across two lanes or two instilments, the other 26 samples were sequenced in a single lane each. This resulted in 57 pairs (forward and reverse) of sequence files.

2.4 Reference genome preparation

The reference genome used was provided by the authors of Glanzmann et al. 2016 [7]. It is a high coverage (~x90) Cape buffalo (*Syncerus caffer caffer*) genome assembly consisting of 1 235 scaffolds which span 97.9% of the estimated 2.732 Gigabase (Gb) length of the animal's genome. Since the assembly is on the scaffold level and not assembled into chromosomes, the total 442 402 scaffold and contigs were joined into 51 super-scaffolds. These super scaffolds were used as artificial chromosomes to ensure reference genome compatibility with most processing and analytical pipelines.

In order to crate these super-scaffolds, ScaffoldStitcher (version 1.0)[72] software was used. Scaffolds and contigs were joined separately since ScaffoldStitcher cannot combine more than one sequence type at a time. First, scaffolds were merged with a maximum length of 60 megabases (Mb), after which contigs were combined with a maximum length of 100Mb. To prevent reads from mapping to the same super scaffold, 1000 nucleotide (N) spacers were used for both scaffold and contigs. The scaffold and contigs combined into 45 and 6 super-scaffolds, respectively. The 51 super-scaffolds (Super_Scaffold0 - Super_Scaffold50) closely emulated the diploid (2n) number of chromosomes of the African buffalo (52).

2.5 Sequence quality control and pre-processing

Before the whole genome sequences were processed, basic quality control was performed on the raw reads. This was to ensure that the processed data is of high quality and that results produced from the data is as reliable as possible. The raw sequences were provided in the fastq file format. This file format is a derivative of the basic fasta file

format, which contains the raw sequence reads, as well as quality data for the reads. FastQC (version 0.11.5) [73] was used to perform a quality analysis on each fastq file. FastQC produces a report in html format for each fastq file which contains the quality summary of the reads/sequence. The summary contains several categories in which the quality metrics are displayed. It will be indicated for each category whether the sequence has passed, failed or warning based on its quality. Once all the fastq files were analyzed with FastQC, MultiQC [74] was used to combine the results in the report files into a single summary file. The resulting report file was used to obtain an overview of the quality of all the sequence data at once.

After the sequences has passed quality checks, all the sequences from samples that were sequenced across two lanes or sequencing instruments were combined into individual files. This was done using the '*cat'* command in bash to combine the fastq files. For every sample with split sequences the two forward fastq files were combined, and the two reverse fastq files were also combined. This resulted in two fastq files per sample (one forward and one reverse). A total number 40 paired sequenced were used (1 pair per individual sample). This equates to a total of 80 fastq files.

2.6 Whole genome sequence processing

2.6.1 Sequence alignment

The first step in processing the genome data is to align the sequence reads to a reference genome. To do this, the Burrows-Wheeler Aligner (BWA) (version 0.7.17) [75,76] software was used. The scaffold-stitched reference genome that was prepared was first indexed using the '*bwa index*' command in BWA with the '*-a bwtsw*' option. Thereafter, the sequences were aligned to the 51 super-scaffold reference genome. The BWA-MEM

algorithm was used to perform the alignment. The commands were excecated in the algorithm's paired end mode, with default setting other than the '*-M*' option enabled, which marks split reads as secondary. This was done to make the output compatible with Picard tools, such as '*MarkDuplicates*'. The input for BWA-MEM are two paired fastq (forward and reverse) files, and the output is a single file in Sequence Alignment/Map (SAM) format. The SAM files were then converted to Binary Alignment/Map (BAM) files using SAMtools (version 1.9)[77] with the '*view*' command with the additional '*-b*' and '*-S*' options enabled. The '*-b*' option sets the output file format to BAM, and the '*-S*' option prevents compatibility conflicts with the input SAM files in some instances when enabled. The BAM files were then validated using Picard (version 2.20.3)[78] '*ValidateSamFile*' function. The validation reported that all the reads in each BAM file were missing read groups, which are required for the Picard '*MarkDuplicates*'. To resolve this, Picard's '*AddOrReplaceReadGroups*' function was used to add the read group information for each file that was obtained from the original file names of the raw fastq files as they are generated by the Illumina sequencing instrument. The read-group information added using the command included the read-group ID (*'--RGID'*), read-group library (*'--RGLB'*), read-group platform (*'--RGPL'*), read-group platform unit (*'--RGPU'*) and read group sample name (*'--RGSM'*). The BAM files were then sorted by coordinates using the '*sort*' function in SAMtools, specifying the output format as BAM with the '*-O bam*' option. The sorted BAM files were subsequently indexed using the SAMtools '*index*' command and then validated again using Picard '*ValidateSamFile*'. This result was one BAM file per sample, for a total of 40 BAM files.

2.6.2 Duplicate removal

Following the validation, the BAM files were used as is for further downstream processing and analysis. Additionally, another identical set of 40 sample BAM files were first processed using Picard '*MarkDuplicates*' to remove duplicates reads (with the '*REMOVE_DUPLICATES*' option set to 'true'). Duplicate reads can cause the quality of called single nucleotide polymorphism (SNP) to be inflated due to overestimation, however the relative effect of this could be interpreted as negligible [79,80]. There are two types of duplicate reads, namely optical duplicates and PCR duplicates. The latter is the result of biased PCR amplification during sample library construction, and optical duplicates caused by the sequencing instrument's optical sensors erroneously detecting multiple clusters instead of single amplification clusters. Basic statistics were generated for all BAM files using SAMtools *'-stats*' and *'-coverage'*.

2.6.2 Genotype likelihood calculation

Next genotype likelihood calculation and variant calling was performed, using BCFtools (version 1.6.33) [75,77]. BCFtools '*mpileup*' command was used to produce genotype likelihoods from the input BAM files. The resulting output file was directly passed ('*piped*') to BCFtools '*call*' for variant calling. Two methods were used for the '*mpileup*' step: The first used information from only a single sample in BAM format to generate the genotype likelihoods; the other used information from all the samples' BAM files to perform a simultaneous likelihood calculation for each of the 51 super-scaffolds. The data produced from the iteration demonstrated that the single sample method produced data that was of insufficient quality for most of the analyses. Ideally, the multiple pileup method would be used to generate the genotype likelihoods across the entire genome for all samples simultaneously, however the large amount of data meant that this would not be practical.

Thus, it was decided to split the process to be performed on each super-scaffold individually, as one would using chromosomes, to increase the efficiency of the process [81].

2.6.3 Variant calling

Both methods were performed using default settings with one notable exception. For the latter method the '*-r*' option was used to define the genomic region as a single super-scaffold for each iteration of the '*mpileup*' steps. In all instances, the BCFtools '*call*' function was executed in the multi-allelic mode ('*-m*' option) and called only variant sites using the '*-v*' option. The final output was Variant Call Format (VCF) files containing variant information for each individual sample or a single super-scaffold across all samples. The VCF files were then compressed to vcf.gz file format using the '*bgzip*' tool included in the SAMtools package. This was done as various tools require compressed and indexed VCF files as inputs. The compressed VCFs were subsequently indexed using BCFtools '*index*'.

2.6.4 VCF files and merging

Three main processing pipelines were used for the creation of variant information in the form of VCF files from the 40 samples' whole genome sequences. The first used BAM files which did not have duplicate reads removed and used the individual genotyping method, also known as Individual Genotyping (IG). The second also did not have duplicates removed but used the multiple pileup method of genotyping (Combined Genotyping (CG)). The third had duplicates removed and used the latter method of genotyping (CG with duplicates removed (CGDR)) (Figure 2.2). All three produced VCFs that required combining to analyse all the samples' variant information simultaneously

and efficiently. In other words, three VCFs were produced containing variant information from all 40 sample genomes, one for each output batch (IG, CG and CGDR). For the IG VCFs, the BCFtools '*merge*' function was used to combine the files. The function is designed to combine VCF files containing information from single samples (one sample per file). The tool used for combining the CG and CGDR VCFs was Picard's '*GatherVcfs*', which is designed to combine VCF from multiple samples and genomic regions. All the combined VCFs were compressed to vcf.gz format and indexed using the same method previously mentioned.

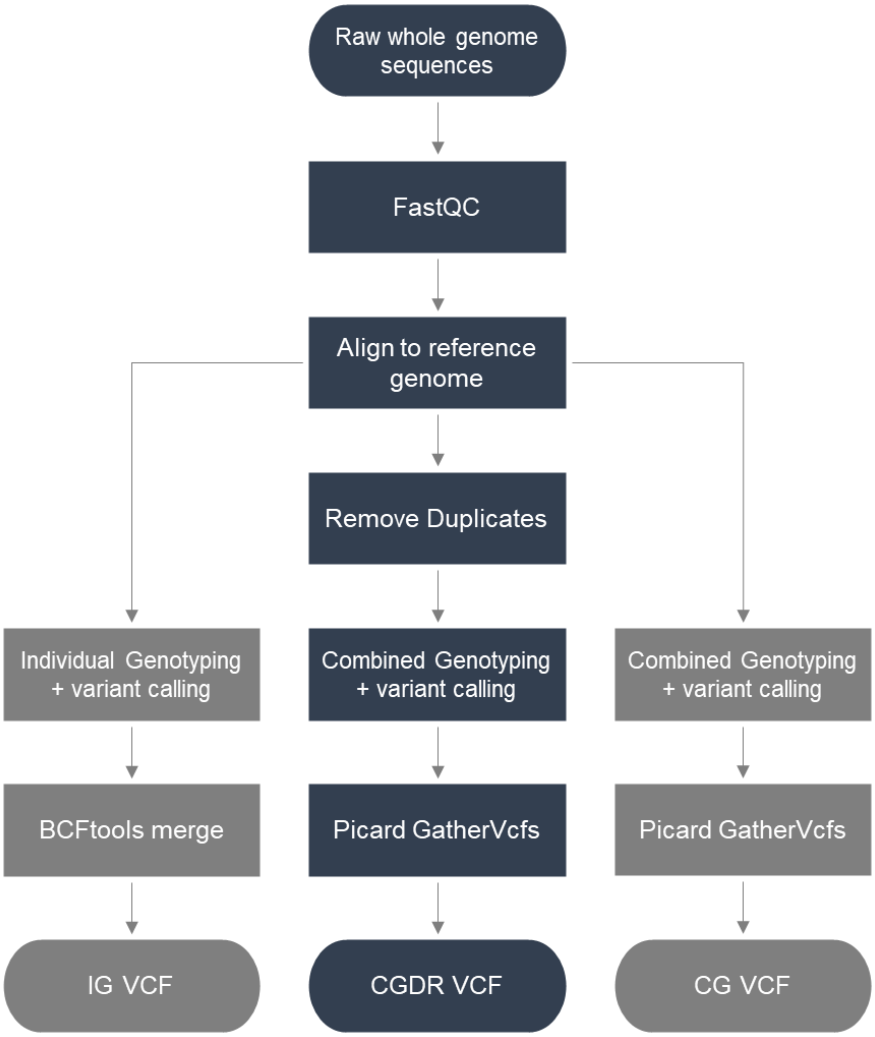


Figure 2.2: A flow diagram depicting the different processes used to obtain the tree VCF files. Note that the centre process was deemed to be the primary processing pipeline.

2.7 Basic variant filtering

The variants in the combined VCF files needed to be filtered for basic quality and variant type. BCFtools, combined with another software package, VCFtools (version 0.1.17) [82] were used to filter the data. Only SNP information would be used for subsequent

analyses, thus all non-SNP variant such as insertions/deletions (INDELs) were filtered out. Several gene-based association tests only use SNPs, due to the confounding effect INDELs and other variant types can have on the statistical identification of associations. Several tools used in this study also only use SNP data to generate population statistics. Similarly, only biallelic SNPs were selected for further analysis. The other quality metrics were selected with standard measures uses for SNP filtering for use in analyses performed in this study.

The VCF file was first processed by BCFtools using the '*view*' command, which passed its output directly to VCFtools for further filtration. BCFtools was set to include only biallelic sites, using the '*-m2'* and '*-M2*' options to set the minimum and maximum number of alleles per site to 2, respectively and the '*-v*' option was set to '*snps*' to include only SNPs. All further filters using BCFtools were classified under the '*-i*' option, which indicates to the command to only include sites which match the criteria listed. Only variants with a minimum quality score of 30, with a minimum site depth and minor allele count (MAC) of 2 were included as well as all variants with a total site depth and allele frequency of less than 500 and 0.9, respectively. Sites with excessively high depth and allele frequencies are generally considered to be sequencing errors which can induce false positives [80]. The filtered variants were once again output as a compressed vcf.gz file which was subsequently index. Basic variant statistics were generated using the BCFtools 'stats' command before and after the filtering step.

2.8 Population statistics

2.8.1 Summary

The structure and genetic background of a population can have a major influence on genome wide genetic analysis. This is especially true for an association test where factors such as inbreeding and population stratification can confound any association found if not accounted for correctly. Various population statistics can also aid in the quality control of variants to be used for said association analyses. The way the information gathered from these statistics is context specific, with implementations varying based on the goals of the study in question. There is a rough consensus in literature on the way in which parameters should be handled [83–88], through either filtration or adjusting subsequent analyses to mitigate or account for the parameters. Two of the key population statistics to take into consideration when performing association studies, are population stratification and the relatedness/heterozygosity of the samples. There are several methods of generating such statistics, and each can have different outcomes. Thus, multiple methods were used to compare their results so that the most effective way of conducting future analyses can be determined. VCFtools and PLINK (version 1.9) [89,90] were the software packages selected for this, as they can perform similar functions using different algorithms/parameters. Both tools use the variant information within a VCF or equivalent file to generate their results. The methods were used on all three VCF, with the IG VCF being the first to be processed, the results of which were used to inform the decisions that led to the usage of the CG and CGDR groups.

2.8.2 Inbreeding

To measure the heterozygosity of the samples, the inbreeding coefficient (F) was estimated by both tools using a method of moments. For VCFtools, the '*--het*' function was used using default settings and stating the input file and its format as vcf.gz with the '*--gzvcf*' argument. The input files were unaltered for use with VCFtools. PLINK was also instructed to use the VCFs as input with '*--vcf*'. However, PLINK requires a linkage disequilibrium (LD) pruned input file to perform its analysis, as the function does not take LD into account. This was the case for all PLINK analyses performed in this study, thus the input files' variants were appropriately LD pruned before each PLINK analysis using the PLINK '*--indep-pairwise*' function/argument. The parameters given to the argument set the window size to 50 kilobases (kb), the step size to 10, and the r^2 threshold to 0.1. Additionally, for every PLINK command used in this study the following options were used: generate missing variant IDs ('*--set-missing-var-ids*'), allow for additional chromosomes ('*--allow-extra-chr*') and set the family IDs to sample IDs ('*--double-id*'). This was done to accommodate the non-standard/non-human data used as input, as PLINK expects pedigree information and the human number of 23 chromosomes. The prune command's output includes two files containing the list of variant IDs which passed and failed the pruning filter. This would then be used to indicate the sites which should be kept in the subsequent PLINK analysis using the '*--extract*' option. PLINK's inbreeding command ('*--het*') in its default configuration uses imputed minor allele frequencies (MAF(s)), this can be changed to use a file containing MAF information. The PLINK inbreeding analysis was performed in 3 separate configurations: default, with a predetermined MAF file ('*--read-freq*') and with the '--small-sample' option enabled. It should be noted that the PLINK method is not necessarily appropriate for sample sizes

as small as the cohort used in this study, thus the results are likely to be reflective of this and thus serves as a secondary result. Inbreeding results from both the PLINK and VCFtools were processed and plotted in R (version 4.0.1) [91,92].

2.8.3 Population structure

A principal component analysis (PCA) plot was used to investigate the population structure within the cohort. A PCA uses the variant information from all the samples to calculate principal components which in turn, act as axes of variation in the samples. These principal components are then ranked based on the proportion of variation it represents in descending order. For the PCA plot, the two components which account for the highest and second highest amount of variation serve as the main axes of the plot. Samples will cluster on the plot based on their relation to one another, in other words the samples will group according to population. This thus serves as an indicator for population structure. PLINK was used to perform the PCA on the samples, using all the prerequisites stated previously, such as LD pruning. The '*--pca*' function was used to instruct PLINK to perform the analysis which produces eigenvalues and eigenvectors as the output. The output was then transferred to an R script, which was then used to make the PCA plot.

2.8.4 Additional statistics

Following the two prior analyses, it was determined that only the CGDR VCF file would be used for all further methods, due to it theoretically being the highest quality of the three sets. Further, more statistics were generated using the VCF files to better understand which filter criteria should be used to generate the SNPs that would be selected from for the candidate variant selection. Statistics indicating the missingness of genotypic information is an example of a statistic that can aid in reducing erroneous variant.

Missingness was calculated on both an individual and site bases using PLINK's '*--missingness*' function. VCFtools was also used to do this however, the output size was such that R was unable to process it, consequently those results were not reported. This was also the case for the subsequent statistical processes involving VCFtools. The other two statistics generated from the VCF were the MAFs, and the Hardy-Weinberg equilibrium proportion p-values for all sites. The PLINK command used for these were '*--freq*' and '*--hardy*', respectively, of which the outputs were plotted using R with in-house scripts.

2.8.4 Power calculation and MAF filter parameter creation

Following these statistical procedures, the information gathered was used to determine the metrics by which the variants would be filtered in tandem with the recommendations from literature on related subjects [84,88,93–95]. In addition to this information, a statistical test was performed to determine the minimum MAF a candidate variant would require, to give a subsequent genetic association test for bTB susceptibility in African buffalo enough statistical power. The web based Genetic Association Study (GAS) Power Calculator [96] was used with parameters derived from the two studies performed by Le Roux et al. [11,63], as the subsequent association study would invariably be performed using the samples from study cohort used in said studies. Thus, the number of cases were set to 200 (actual number was 198 individuals [63], however the tool automatically rounds up to 200) and controls were set to 670 [63]. The significance threshold (α) value was set to 0.05 and the disease model was set to additive, as to stay in line with the previous study [63]. The average reported prevalence of bTB with in the two populations (KNP and HiP) of the cohort was 23% [63], therefore the corresponding parameter was set to 23% for the test.

The genotypic relative risk was set to 1.5, implying that individuals with the associated disease variant would have a 50% grated risk of developing the disease. The disease allele frequency was then adjusted until the power of the test was 0.80, which would serve as the cut-off for having enough power for a one-stage association study.

2.9 SNP panel creation

2.9.1 Variant filtration

After all the necessary information was obtained, the final filter was prepared and implemented with the following parameters: Minimum MAF: 0.10 (BCFtools '*filter -i* '*MAF>=0.10*'); Maximum missing site genotypic data: 5% and minimum Hardy-Weinberg equilibrium (HWE) p-value: 1×10e-6 (VCFtools *'--max-missing 0.95 --hwe 0.000001'*). Using the commands given here the variants in the VCF were filtered and the new VCF file was compressed and indexed for further use.

2.9.2 Gene panel

A total of 26 genes were selected for the panel, all of which are immune related genes which have been previously associated with TB infection, primarily in humans [56,63]. The genes' sequence files of the cow homolog used for the panel are represented in Table 2.1. These were obtained as a part of the cow reference genome assembly ARS-UCD1.2 (GCF_002263795.1). The genomic regions containing the buffalo counterparts within the modified reference genome were located using Basic Local Alignment Search Tool nucleotide (BLASTn) [97]. The command line version of the software, BLAST+ (version 2.7.0+) [98], was first used to create a local database for the scaffold-stitched buffalo reference file with the '*makeblastdb*' with the option '*-dbtype nucl'* to indicate that it should be a nucleotide database. The gene sequences were all processed as the query

sequence ('*-query*'), through BLAST+ against the new buffalo local reference database individually, using the 'blastn' command. The resulting alignments were then used to restrict the variants in the filtered VCF to only sites which fall within the corresponding gene regions.

Table 2.1: List immune related cow homolog genes previously associated with TB infection from the cow reference genome selected for the SNP panel.

Genes*	NCBI reference**
BOLA-DRB3	NC_037350.1:25723691-25734819
CCL2	NC_037346.1:c15905265-15903398
CD209	NC_037334.1:c16589774-16586871
CXCL10	NC_037333.1:c90884345-90881993
CXCL8(IL8)	NC_037333.1:88810807-88814572
DMBT1***	NC_037353.1:42390305-42492218
IFNG	NC_037332.1:45624513-45629336
IFNGR1	NC_037336.1:c75106378-75081984
IL10	NC_037343.1:c4555309-4551373
IL12B	NC_037334.1:c70911817-70890911
IL1A***	NC_037338.1:c46493533-46482553
IL1B	NC_037338.1:c46551922-46543408
MBL2	NC_037353.1:6331933-6337539
MC3R	NC_037340.1:c59648829-59645909
NOS2(NOS2A)	NC_037346.1:c19431368-19388579
P2RX7	NC_037344.1:c54027496-53966935
PTPN22	NC_037330.1:29519848-29578366
SFTPA1	NC_037355.1:c35649951-35645844
SLC11A1(NRAMP1)	NC_037329.1:106392721-106403646
SLC7A13***	NC_037341.1:76487564-76495653
SP110	NC_037329.1:c118232904-118183281
TLR2	NC_037344.1:c3990089-3953807
TLR4	NC_037335.1:107057826-107068842
TLR8	NC_037357.1:c130745827-130728235
TNF	NC_037350.1:c27718943-27716170
VDR	NC_037332.1:32333822-32441112

* gene names represent the genes in the cow, and Human equivalent in brackets where applicable
** for the cow reference genome genome ARS-UCD1.2 (GCF_002263795.1)
** *Genes that were found to be associated with bTB susceptibility in African buffalo

2.9.3 Association test panel

To test performance of the SNPs in the panel in an association test, a test was performed using available data. To narrow down the search for candidate variants as a matter of practicality, three of the 23 genes in the panel, were selected to focus the association test on likely candidate genes/genomic regions. Two of the genes, interleukin 1 alpha (*IL1α*) and solute carrier family 7 member 13 (*SLCA13*) were both previously associated with bTB susceptibility in African buffalo in the study by Le Roux et al. [63] The third gene *BOLA-DRB3*, represents a homologous proxy to the human major histocompatibility complex (MHC), which is a major driver of immunity and susceptibility [99–101]. The alignment files were then used to correlate and convert the SNP coordinates from the buffalo reference genome to that of the cow genome ARS-UCD1.2 (GCF_002263795.1). This was done by making use of the Ensembl Variant Effect Predictor (VEP) [97] online web interface to determine the consequence and general characteristics of the SNPs within the panel. In summation, the narrowed down SNP panel serves as the basis for the subsequent test of the panel in a targeted genetic association test.

2.10 Targeted association test

The test panel consists of the 143 variants from the 3 selected genes, and regardless of the VEP results, as the possibility of all variants acting as markers for true associated variants/genes. This was performed as a simple targeted case/control association test which would include all 40 samples used thus far. The bTB exposed sample (from KNP and HiP) would serve as the cases (n=30), while the unexposed AENP and MNP samples would be the controls (n=10). First as part of the analysis a Fisher's exact test was performed using PLINK '*--assoc fisher*' function, and the command was set to output 95%

confidence intervals for the odds-ratio (OR) with the '*--ci 0.95*' option. Thereafter a Cochran-Armitage trend test ('*--model trend*') was performed for an association test and was seen as the most appropriate model (as it was also used by Le Roux et al. 2013), as it does not assume HWE is in effect. Thus, it fits the population stratified nature of the samples used. There was also a logistic regression model test performed using the *'--logistic'* option. Additionally, to account for population structure, the output eigenvector file from the CGDR PCA was used as covariates for the association test using the '*--covar*' option as well as the loading of a cluster file (using the '--cluster' option) which assigns each sample to a cluster. From the PCA results it was determined that there were three distinct cluster for the samples. The PLINK association commands were performed using all the basic parameters and options as previously described with the exception of linkage pruning which was not performed on the SNP panel VCF file. The small sample size and the skewed cases/controls ratio would mean that the test would be severely underpowered, however the purpose of the test was not to produce significant results, but rather just to test how the SNP panel and adjacent data would perform. However, for the sake of the test a p-value of below 0.05 was considered to be a 'significant' association.

CHAPTER 3:

RESULTS

Contents

3.1 Whole genome sequencing and processing

The whole genome sequencing produced an average of 30.55 Gb of raw paired sequence per sample. The mean total sequence per sample was 71.5 Mb, with the mean GC content and duplicates across all samples was 41.72 % and 15.11%, respectively (Figure 3.2). The MultiQC report generated from the FastQC results showed that all the raw sequenced passed overall quality control (Figure 3.1). Warnings were produced in certain categories, however, these warnings did not warrant any remediation due to the criteria the warnings were based on [73]. The MultiQC output summary (Figure 3.3) showed that some of the sequences failed in the per tile sequence quality and/or Kmer content categories.

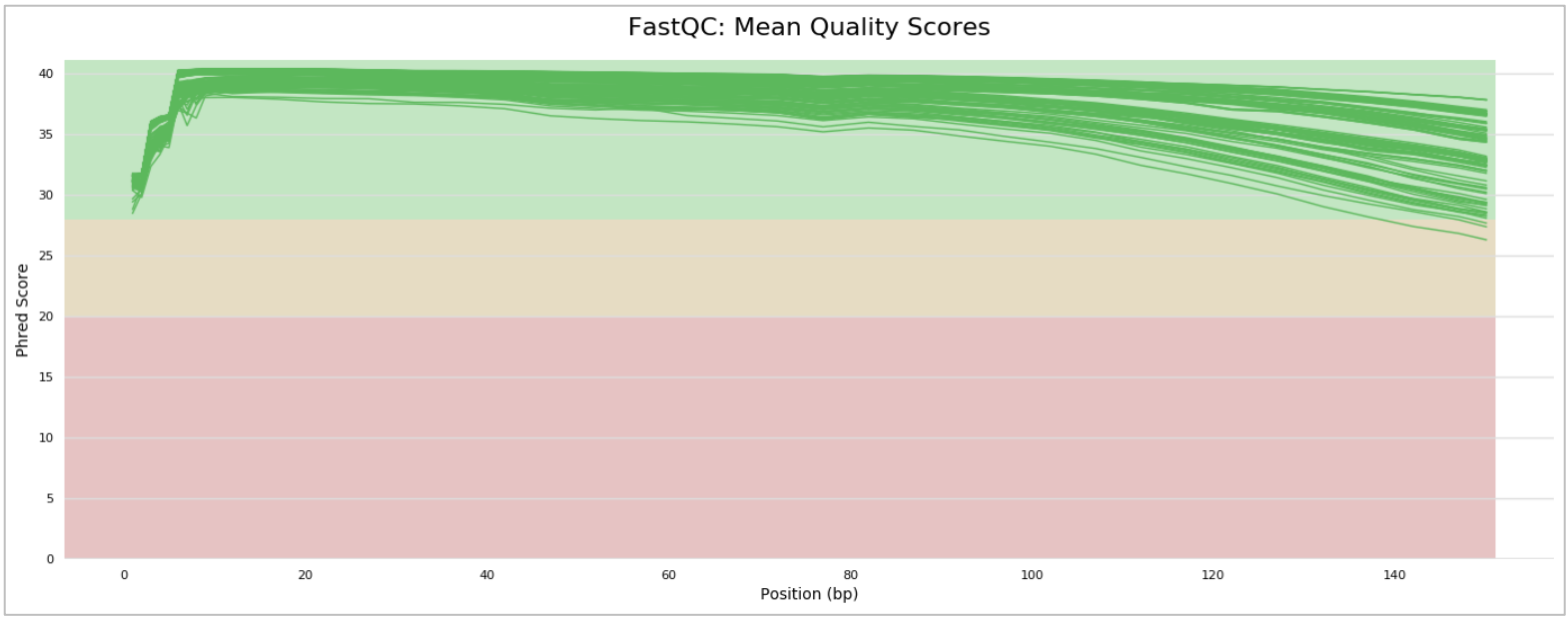


Figure 3.1: MultiQC combined FastQC output for the mean quality value (phred score) across each base position (bp) in the read. Each of the green lines represent a single sequence. Green indicates that the sequence has passed the mean per base quality score of the FastQC quality control test.

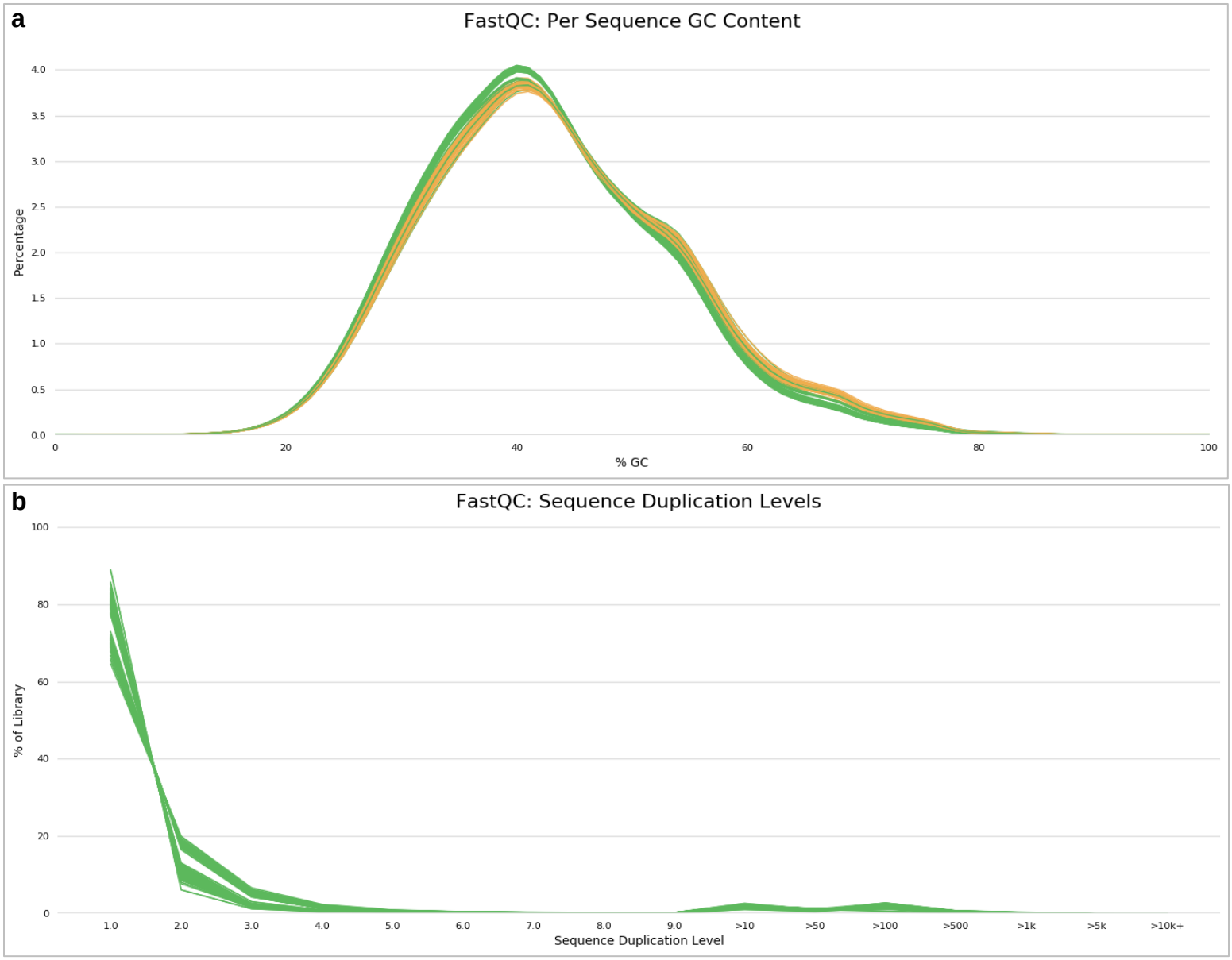


Figure 3.2: MultiQC combined FastQC output for **a)** the mean GC content (% GC) and **b)** mean duplication rate, per read. Green lines represent sequences which passed quality control, and orange lines represent sequences that passed with warnings.

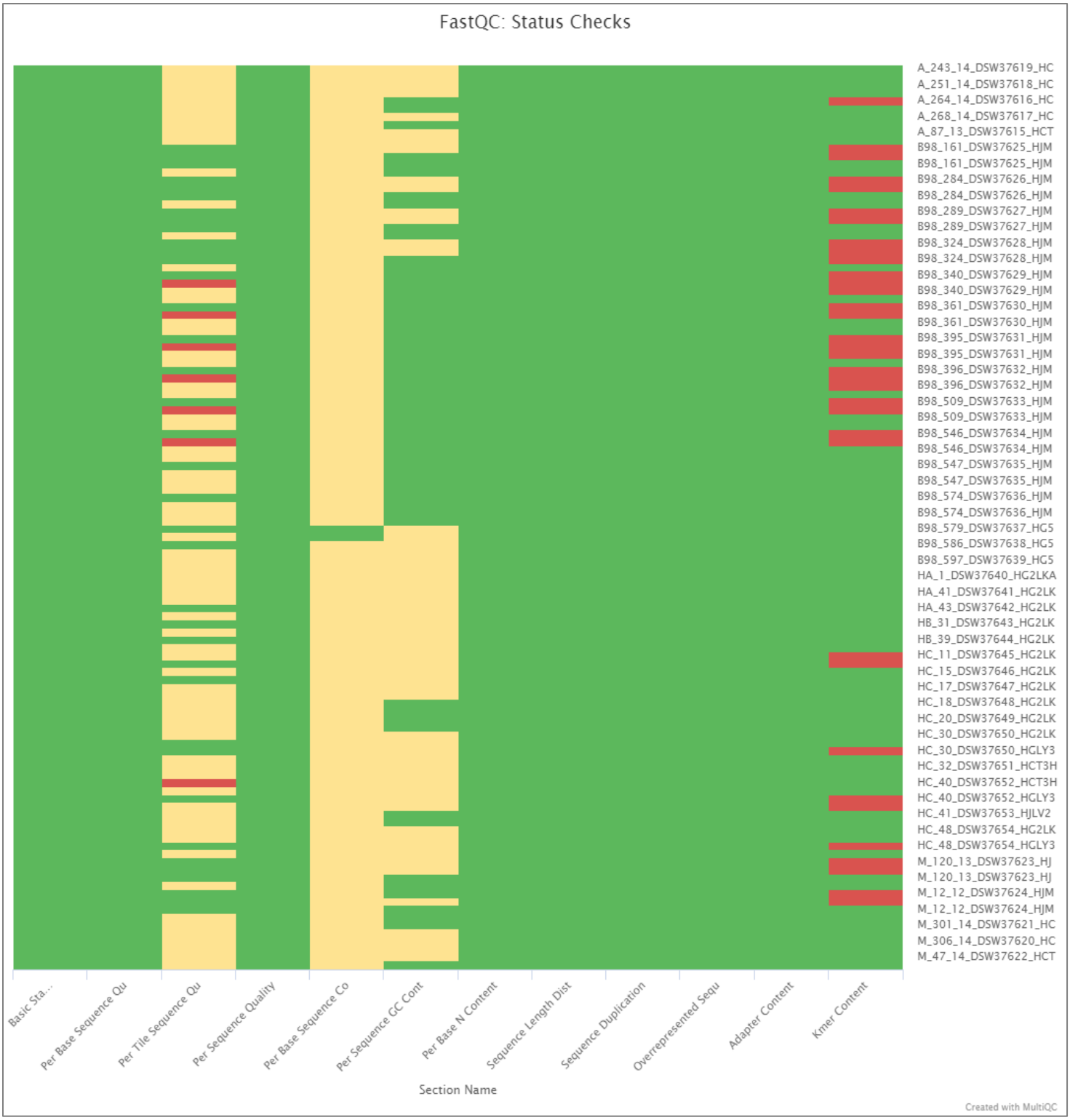


Figure 3.3: MultiQC output depicting the status of each sequence (right) for each FastQC category's (bottom) results shown as: are shown as: "entirely normal (green), slightly abnormal (orange) or very unusual (red)"

The statistics produced after the sequence alignment and duplicate removal are represented in Table 3.1. The mean coverage of the sequence across all samples was 96.46% with and without duplicates present. The mean depth of coverage was 10.01x with duplicates present, while it is lower at 8.28x with duplicates removed. The mean base- quality and mapping quality for both duplicates present and removed were similar (Table 3.1).

Table 3.1: Statistics generated from sequence alignment BAM files using SAMtools.

	Coverage (%)	Mean depth of coverage	Mean base quality score	Mean base mapping quality
Duplicates present	96.46%	10.01×	37.88	53.81
Duplicates removed	96.46%	8.28×	37.01	53.70

3.2 SNP calling and basic quality filtering

The number of single nucleotide polymorphisms (SNPs) called and transitions versus transversions ratio (Ts/Tv) vary slightly between the three processing pipelines (namely IG, CG and CGDR, refer to Figure 2.1), but are similar. The number of SNPs were 50-58 million and the Ts/Tv was 1.89-1.95 (Table 3.2) before filtering. The number of SNP after basic quality filtration decreased for all three pipelines. IG had far less SNPs (~9 million) than CG and CGDR (~36 million and -37 million, respectively) after filtration. There was an increased in Ts/Tv in all 3 VCFs, with IG having a higher ratio (2.10) than CG and CGDR (both 2.05).

Table 3.2: Statistics generated from the 3 VCF from the 3 different processing pipelines files using VCFtools.

	Before filtering		After filtering	
Genotyping method	no. SNPs	Ts/Tv	no. SNPs	Ts/Tv
IG	58 506 465	1.89	9 04 5612	2.10
CG	54 506 827	1.92	36 083 930	2.05
CGDR	50 076 964	1.95	37 270 097	2.05

IG: Individual genotyping;
CG: Combined genotyping;
CGDR: Combined genotyping with duplicates removed;
Ts/Tv: Transitions versus transversions ratio

3.3 Inbreeding

The inbreeding estimates (inbreeding coefficient, F) for each individual for all IG, CG and CGDR from both VCFtools and PLINK are depicted in Figure 3.4. The inbreeding coefficient (F) results were inconsistent between VCFtools and PLINK, with varying degrees of similarly between each iteration of the analysis. The results from the default and MAF-loaded PLINK inbreeding analyses were identical and were thus represented as the same plot, but those only slightly differed from the PLINK outputs with the '*small-sample*' option enabled. All the results for IG were severely negative, which indicated excessively high heterozygosity compared to what is expected. This can be taken as an indication of erroneous or inappropriate genotype likelihoods generated from the IG processing pipeline or a severe contamination event. To investigate this the CG and CGDR processing pipelines were implemented, however the outputs from these pipelines were also inconsistent between the two tools but to a less severe extent for CG. The CGDR results still showed moderately to severe negative inbreeding values for all

individuals for two populations (KNP and MNP) with the exception of one individual (B98_567). When considering all of the CG/CGDR inbreeding outputs, a general consensus can be observed that the AENP and HP populations display the highest levels of inbreeding (~0.1-0.3) and the KNP and MNP individuals display low to no inbreeding apart from B98_567 (F= ~0.1-0.2).

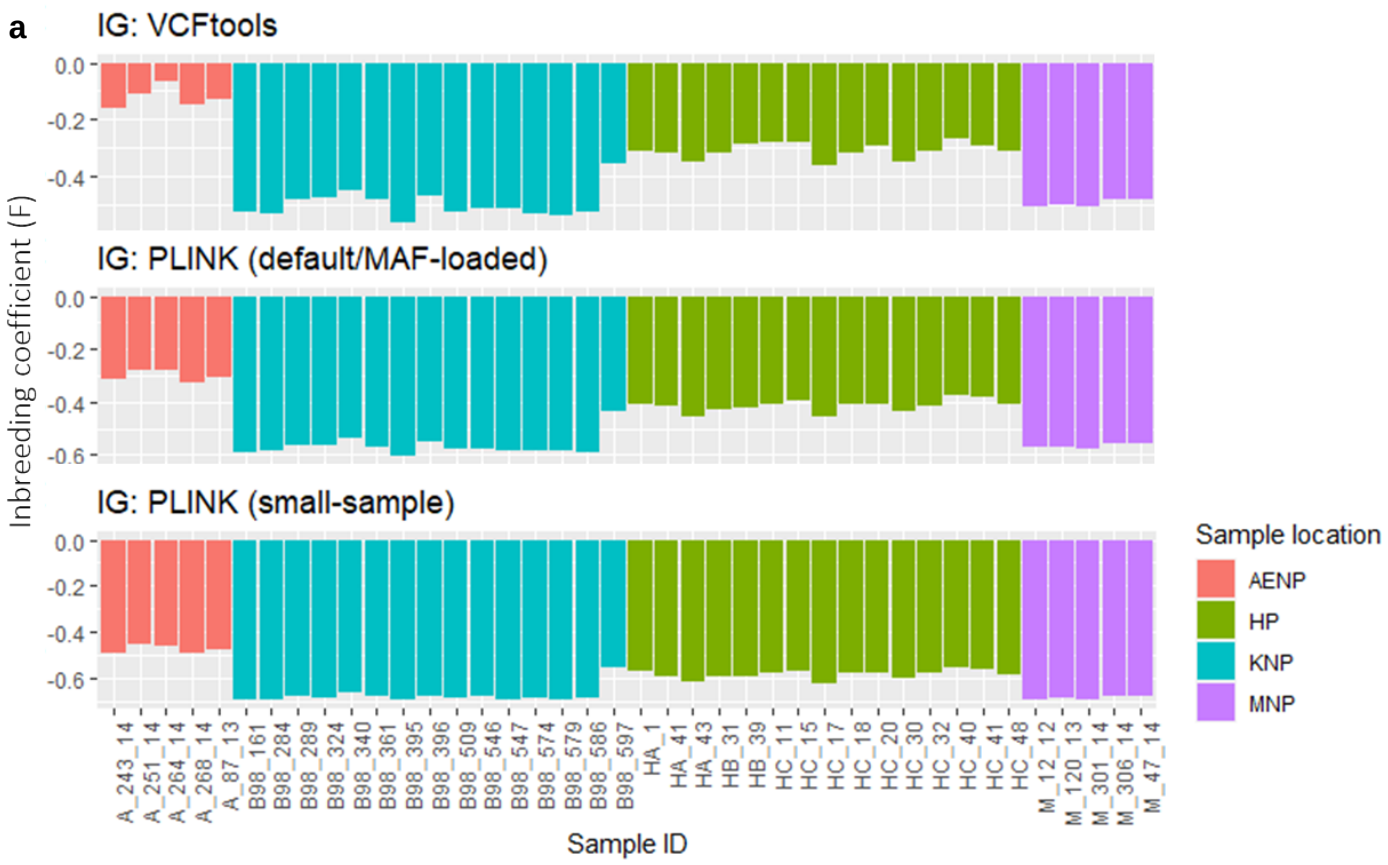
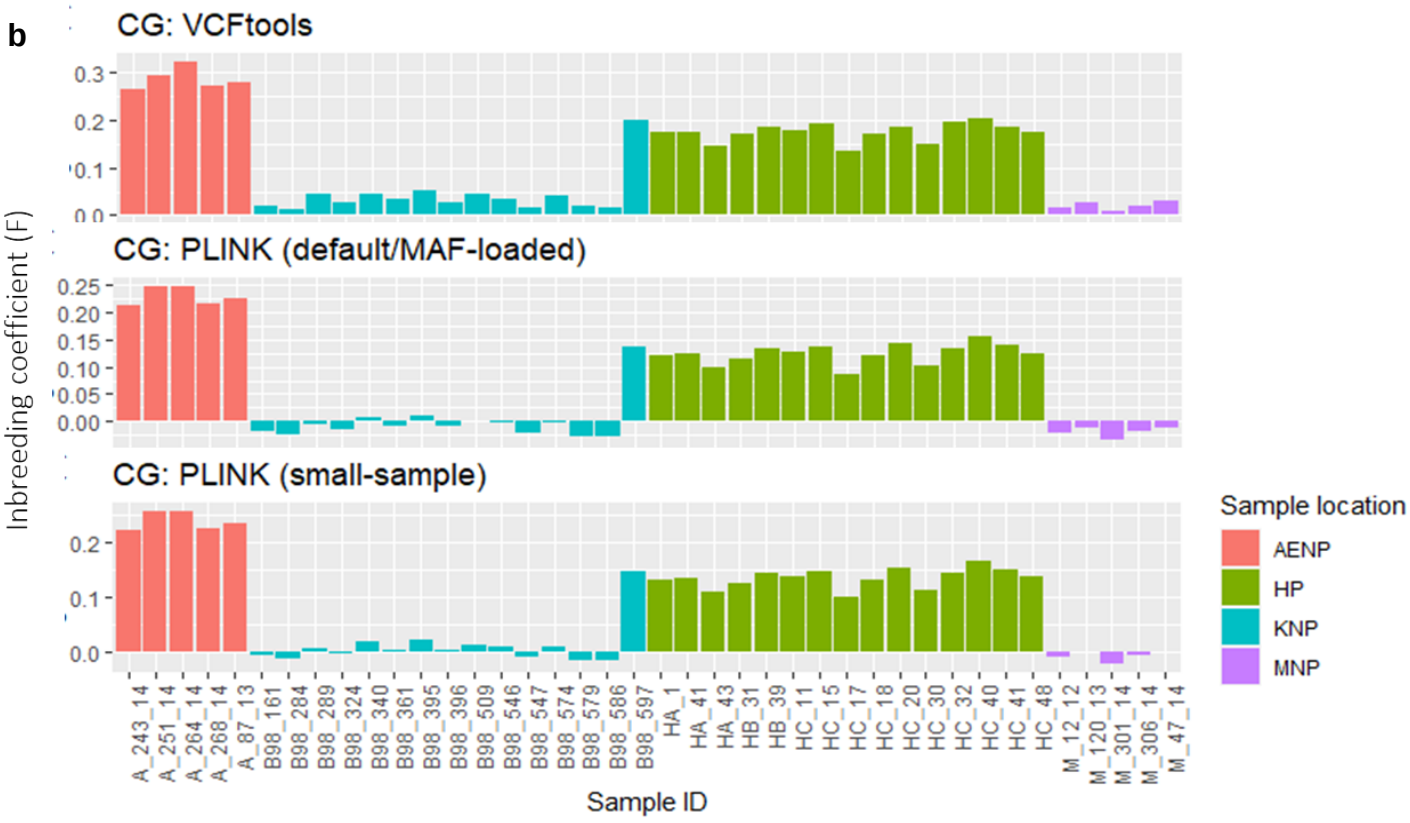
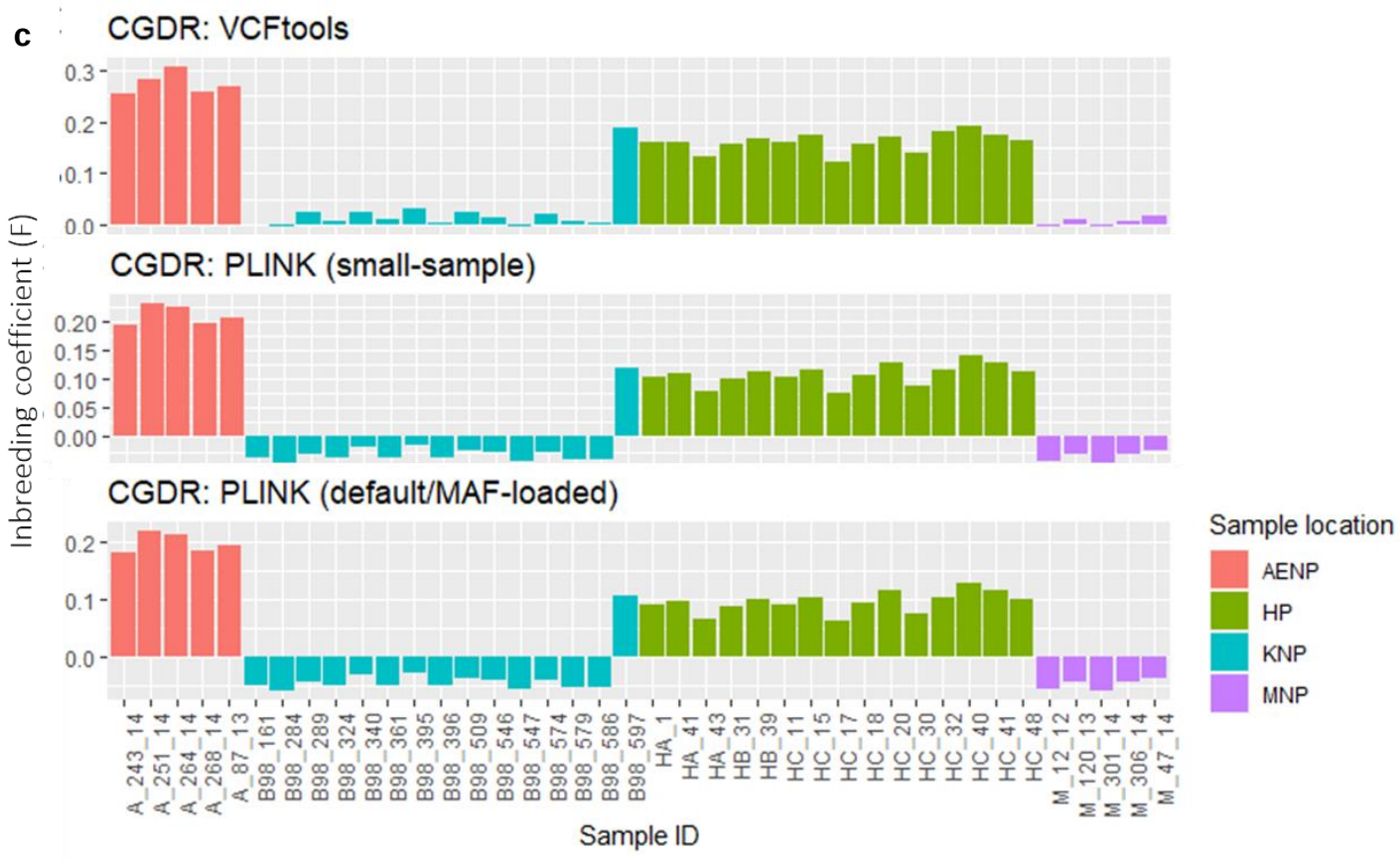


Figure 3.4: Per individual Inbreeding estimates represented as inbreeding coefficients (F) generated using VCFtools and PLINK for **a)** IG, **b)** CG and **c)** CGDR. The PLINK results are split into based on the configuration of the command used. The one being the default or minor allele frequency loaded configuration and the other which used the 'small-sample' option enabled.

[**b)** and **c)** on next page]





3.4 Population Structure

The principal component analysis (PCA) generated by PLINK and plotted using R across all iterations show that the individuals grouped into 3 distinct clusters or populations (Figure 3.5). The individuals in the plot from IG displayed a looser clustering compared to CG and CGDR, and the proportion variance assigned to the principal components (PC) for IG was also higher than CG/CGDR. The plots from CG and CGDR were both nearly identical, only having slightly different percentage variance between the PCs.

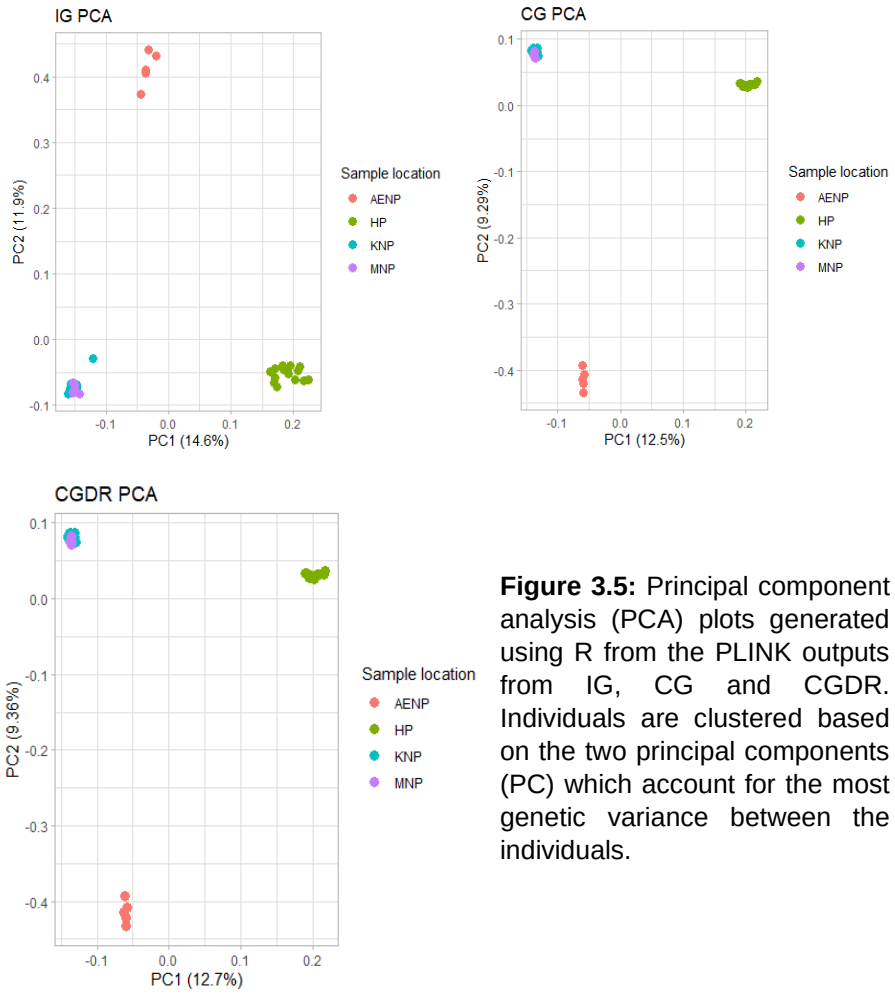


Figure 3.5: Principal component analysis (PCA) plots generated using R from the PLINK outputs from IG, CG and CGDR. Individuals are clustered based on the two principal components (PC) which account for the most genetic variance between the individuals.

3.5 Variant statistics and SNP panel parameters

Variant statistics generated as means of determining filter parameters for quality control for the final SNP panel are depicted in Figure 3.6. These statistics were only produced from the CGDR VCF file as it was the file that would be used to create the SNP panel. The individual genotypic data missingness results showed that no individual had a missingness rate above reasonable levels (Figure 3.6 a). All individuals would pass the recommended general maximum cut off values of 0.05 to 0.25 [84,88], as the highest missing rate was 0.021 which falls well below such filters. Thus, no individuals were excluded based on this statistic. The per site missingness was concentrated around 0.0, where the majority of SNPs were grouped (Figure 3.6 b). This allowed for a stringent per site missingness minimum filter of 0.05. The Hardy-Weinberg equilibrium (HWE) results also produced a distribution were most of the SNPs has high adherence to HWE with -10log adjusted p-values around 0 to 3 (Figure 3.6 c). Considering these results, the discission was made to set the minimum HWE filter a standard 1×10^{-6} [84,88]. According to minor allele frequency (MAF) distribution plot an appropriate minimum MAF filter would be around 0.03 to 0.05 (Figure 3.6 d). However, the MAF filter parameter will be based on the subsequent analysis as outlined in Chapter 2.8.4.

The analysis in question is the power test to determine the minimum MAF needed for variants used in an association test to have enough statistical power given the know sample size and disease (in this case bTB) prevalence within the population. With all the parameters set as described in Chapter 2.8.4., the result was a minimum MAF of 0.10.

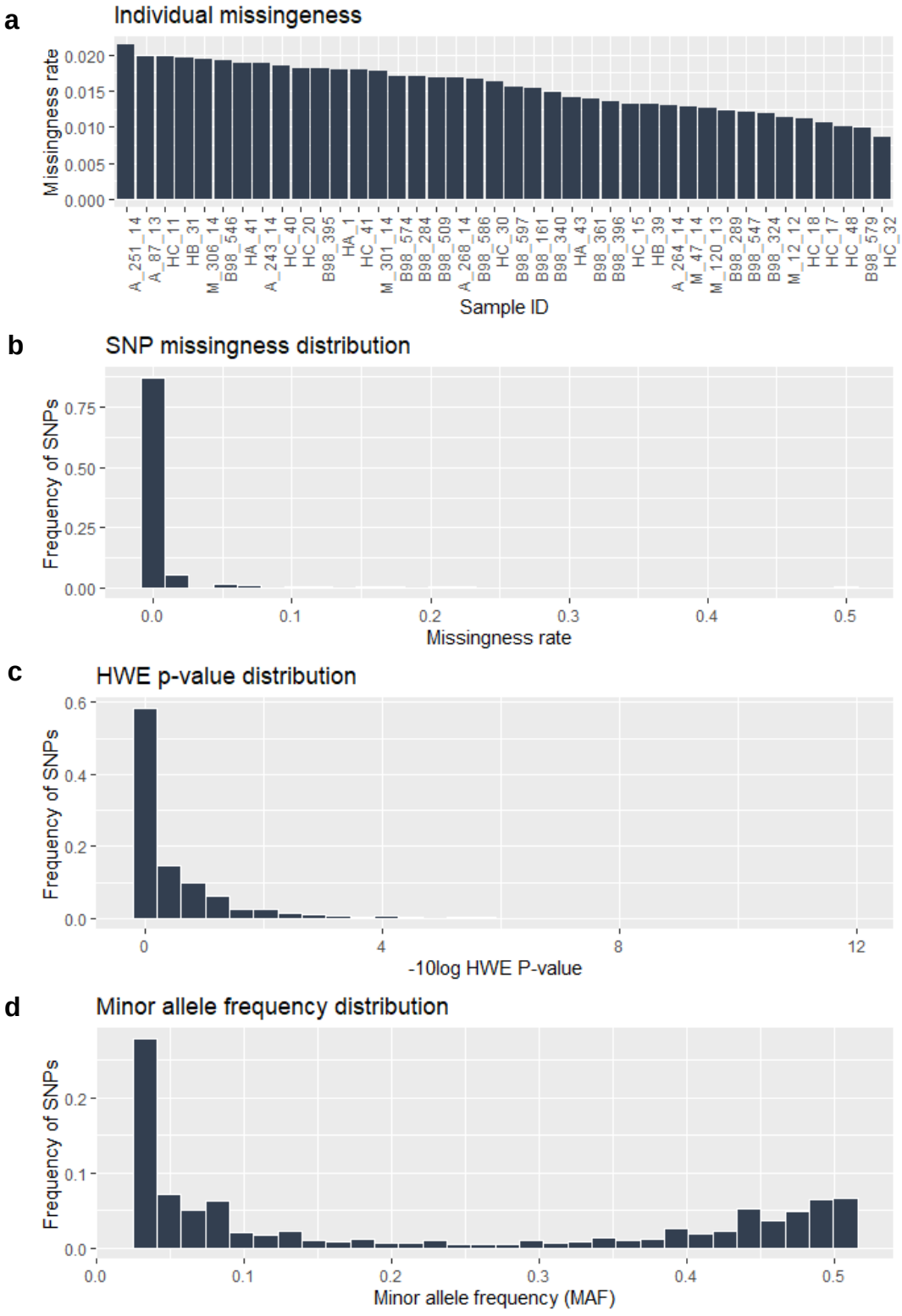


Figure 3.6: Plots depicting the outputs from the individual/variant quality control statistics for: **a)** Per individual missingness; **b)** per site/SNP missingness; **c)** HWE p-value distribution (with -10log adjusted p-values); **d)** MAF distribution

3.6 SNP panel creation

3.6.1 Full 26 gene SNP panel

Once the quality control filter was applied to the CGDR VCF it produced a new file containing just over 20 million SNPs with a Ts/Tv ratio of 2.09. The file was then filtered for the regions which represented the buffalo orthologs of the 26 chosen cow genes and produced a total of 3698 SNPs (Table 3.3). The super-scaffold in which each of the genes were found in the scaffold stitched reference were found and well as the number of SNPs found in each gene, are represented in Table 3.3.

Table 3.3: The list of genes and the super-scaffold in which each were found, and the number of SNPs found in each gene.

Gene*	Super-scaffold	Number of SNPs
BOLA-DRB3	Super-scaffold6	69
CCL2	Super-scaffold12	5
CD209	Super-scaffold21	11
CXCL10	Super-scaffold39	4
CXCL8(IL8)	Super-scaffold22	8
DMBT1	Super-scaffold1	842
IFNG	Super-scaffold38	12
IFNGR1	Super-scaffold37	117
IL10	Super-scaffold26	35
IL12B	Super-scaffold17	153
IL1A	Super-scaffold41	37
IL1B	Super-scaffold41	60
MBL2	Super-scaffold27	79
MC3R	Super-scaffold8	17
NOS2(NOS2A)	Super-scaffold40	319
P2RX7	Super-scaffold0	222
PTPN22	Super-scaffold4	129
SFTPA1	Super-scaffold11	25
SLC11A1(NRAMP1)	Super-scaffold31	65
SLC7A13	Super-scaffold38	37
SP110	Super-scaffold16	277
TLR2	Super-scaffold13	376
TLR4	Super-scaffold21	11
TLR8	Super-scaffold28	54
TNF	Super-scaffold5	10
VDR	Super-scaffold23	724
Total	-	**3698**

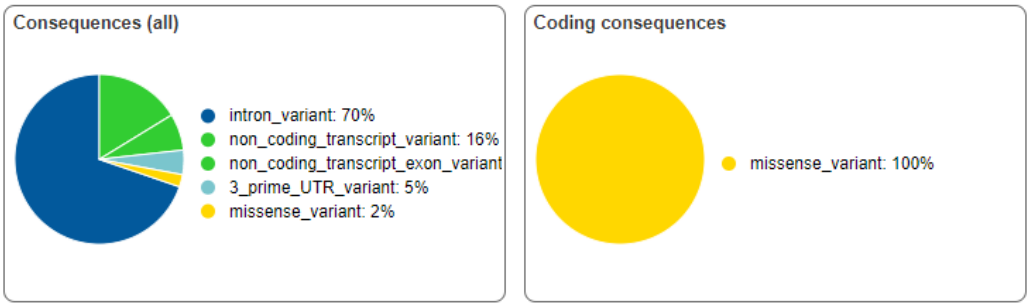
* gene names represent the genes in the cow, and Human equivalent in Brackets where applicable

3.6.2 Association test panel

For the association test subset panel, *IL1α, SLC7A13* and *BOLA-DRB3*, were found by the BLASTn search to be located in Super_Scaffold41, Super_Scaffold38 and Super_Scaffold6, respectively. These regions yielded 37, 69 and 37 SNPs for *IL1α, SLC7A13* and *BOLA-DRB3*, respectively for a total of 143 variants. After these SNPs were isolated from the buffalo data and translated into SNPs with cow genome co-ordinates, the Ensembl Variant effect predictor (VEP) web tool only processed 36 out of the 143 variants (Figure 3.7 a). The SNPs that were excluded by VEP were wildtype in the cow reference genome and thus not present in the output as generated by VEP. The remaining SNPs consisted of 33 novel and 3 existing variants, one of which was the only SNP in a protein coding region. The variant in question (Ensemble ID: rs454104745) is a missense A to G substitution, in exon 3 exon of *BOLA-DRB3*, located on chromosome 23 of the cow genome. It causes an amino acid change of Isoleucine to Valine, with VEP classified impact of moderate, and a SIFT consequence predictive score of 0.12, which means the consequence is considered tolerated. The variant is located in a gene domain which codes for Immunoglobulin C-Type (IGc1) (SMART accession number: SM00407), which involved in several immune related functions. There is no frequency information on the variant, and it has only been reported as being observed twice. The site where the variant occurs is highly conserved when looking at the results from the VEP 95 amniota vertebrates Mercator-Pecan alignment at that region across multiple species (Figure 3.7 b).

a

Category	Count
Variants processed	36
Variants filtered out	0
Novel / existing variants	33 (91.7) / 3 (8.3)
Overlapped genes	3
Overlapped transcripts	3
Overlapped regulatory features	-



b

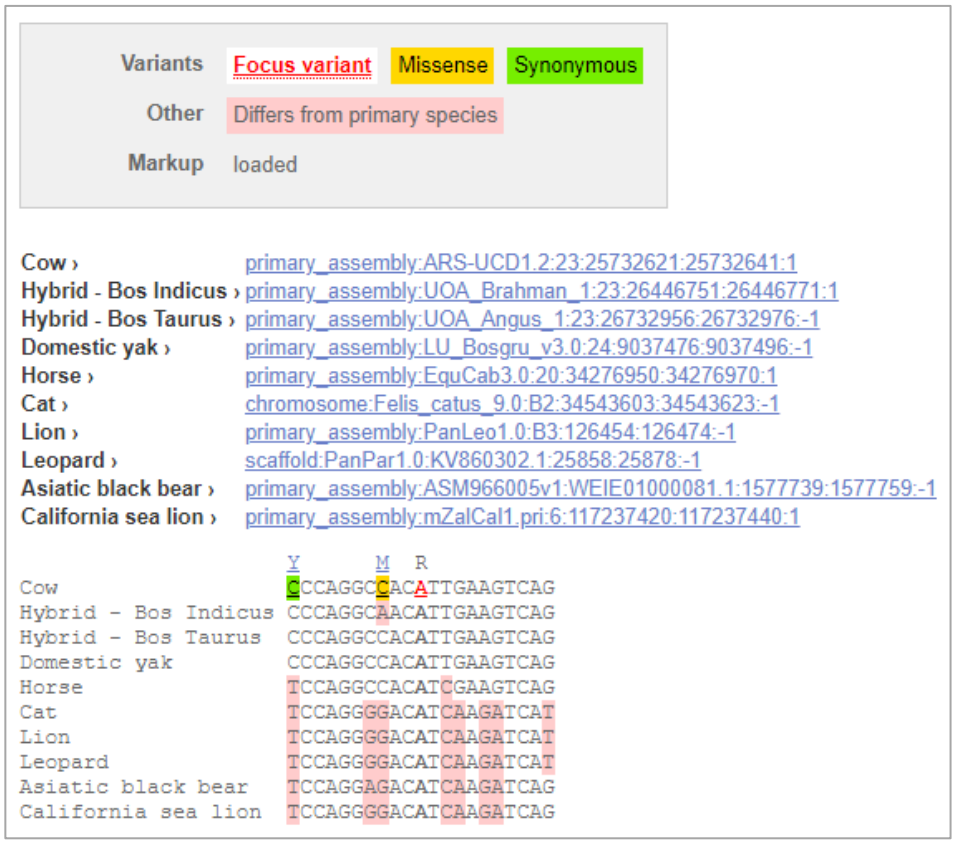


Figure 3.7: Results from the Ensembl Variant effect predictor (VEP) online tool for the final SNP panel. **a)** Basic summary of the amount and proportion of the different types of variants. **b)** The results from the VEP 95 amniota vertebrates Mercator-Pecan alignment at the region surrounding the site of interest (red “A”) where variant rs454104745 (“R” above red “A”) is situated across the species’ genomes which are listed above.

3.7 Association test

The variants which showed significant association with bTB exposure test are represented in Table 3.4, with the odds ratios (OR) and confidence intervals (CI) generated from the Fisher's exact test and the P-values generated by the Cochran-Armitage trend test. The Quantile-Quantile (Q-Q) plot of the trend test p-values (Figure 3.8) showed slight deflation compared to the expected trajectory line, which could be explained by cryptic population covariates which have not been accounted for. However, the general shape and distribution of the data conforms to the sought-after line with flared tale (distant points near the end of the plot) distribution, which indicates reasonably reliable data. The logistic regression test yielded no significant associations (all p-values were effectively 1), thus those results were omitted for the sake of speculative interrogation of the trend test-based associations. There were 10 SNPs which showed significant associations (p-value<0.05), as well as having strong associations indicated by ORs deviating substantially from 1. SNP 120 had the lowest p-value of 7.35×10^{-5}, with an odds ratio of 0.10. The majority (8 out of 10) of the SNPs were in *SLC7A13*, which could point to a trend that the gene in general is more likely of being associated with bTB exposure.

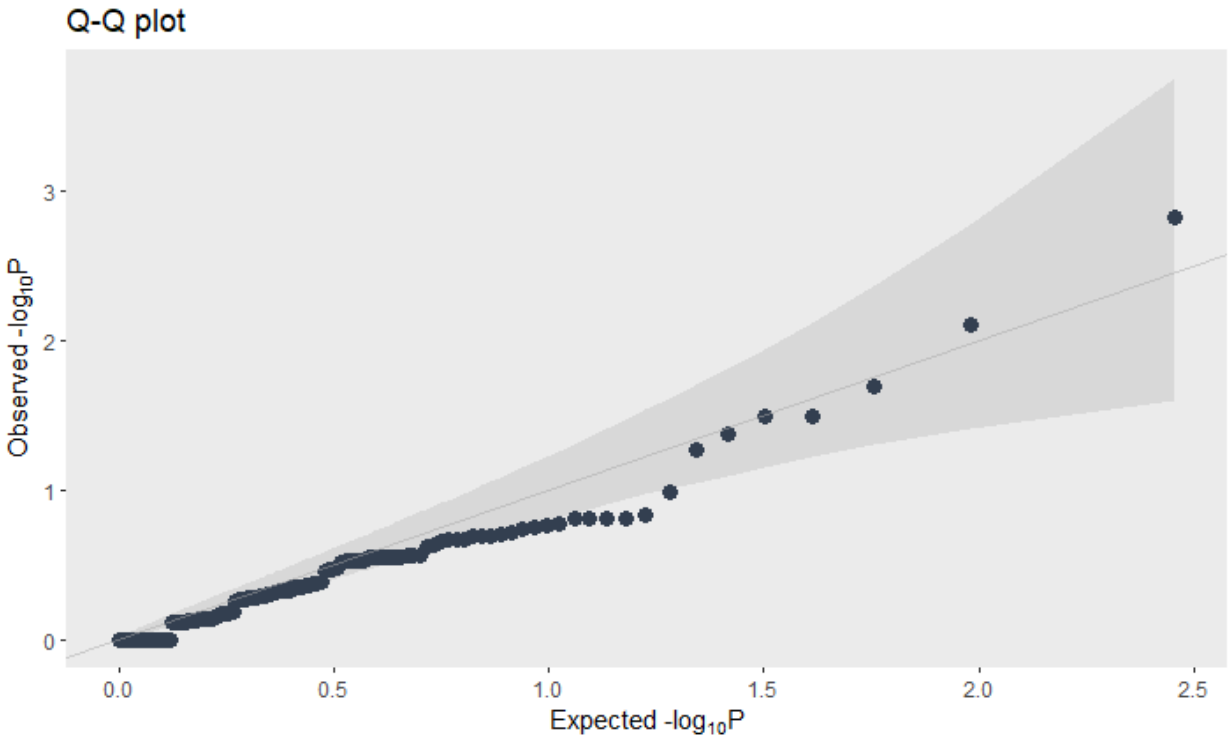


Figure 3.8: The Quantile-Quantile plot generated using the p-values generated from the Cochran-Armitage trend association test as the expected versus the observed p-values. Each point represents a variant, where the points are expected to fall within the darker grey region.

Table 3.4: SNPs from the SNP panel which were significantly associated with bTB exposure based on the p-value generated by the Cochran-Armitage trend test (p-value<0.05)

SNP	Super scaffold	BP	Gene	Allele	FA	FU	SE	L95	U95	OR	p-value
SNP120	Super_Scaffold38	51593334	*SLC7A13*	G	0.07	0.44	0.72	0.02	0.41	0.10	7.35E-05
SNP136	Super_Scaffold38	51597878	*SLC7A13*	A	0.10	0.43	0.69	0.04	0.57	0.15	0.00082
SNP142	Super_Scaffold38	51599005	*SLC7A13*	A	0.13	0.43	0.68	0.05	0.75	0.20	0.00372
SNP118	Super_Scaffold38	51593027	*SLC7A13*	T	0.29	0.05	1.07	0.94	61.62	7.60	0.01006
SNP119	Super_Scaffold38	51593134	*SLC7A13*	A	0.29	0.05	1.07	0.94	61.62	7.60	0.01006
SNP117	Super_Scaffold38	51592929	*SLC7A13*	T	0.28	0.05	1.07	0.90	59.51	7.31	0.01299
SNP111	Super_Scaffold38	51591169	*SLC7A13*	C	0.09	0.31	0.72	0.06	0.91	0.22	0.01558
SNP129	Super_Scaffold38	51596513	*SLC7A13*	G	0.26	0.50	0.66	0.10	1.32	0.36	0.02622
SNP4	Super_Scaffold6	36749161	*BOLA-DRB3*	T	0.25	0.50	0.67	0.09	1.25	0.33	0.02622
SNP27	Super_Scaffold6	36753521	*BOLA-DRB3*	G	0.47	0.25	0.89	0.46	15.15	2.65	0.02840

BP: Base position;
FA: Frequency in cases/affected;
FU: Frequency in controls/unaffected;
SE: Standard error;
L95: 95% confidence interval (CI) lower bound
U95: 95% CI upper bound
OR: Odds ratio;
p-value: The p-value generated by the Cochran-Armitage trend test.

CHAPTER 4:

DISCUSSION

Contents

4.1 Sequence metrics

4.1.1 Whole genome sequencing quality

To produce a panel of high-quality African buffalo SNPs for a targeted association test, high quality whole genome sequences were required. The low coverage whole genome sequence from the 40 Cape buffalo (*S. c. caffer*) individuals that was used in this study was of high quality, with the mean first base quality of above 30 (Figure 3.1). The GC content and genome mean coverage (41.72 % and 96.46% respectively) is highly similar to the African buffalo reference genome produced by Glanzmann et al. which was 41.7% and 97.9%, respectively (Table 3.1). This indicated that the sequences did not deviate from what is expected for high quality sequencing. The Kmer content and per tile sequencing quality failures for some of the raw sequences seen in Figure 3.3 is most likely due to random priming which is common among Illumina and other sequencing platforms [73]. Per tile sequencing content also cannot be remediated at a post sequencing level, thus no action could be taken if indeed it was of concern.

4.1.2 Sequencing alignment quality

The alignments produced from the sequences were also of high quality with a mean phred scaled mapping quality of ~37 as well as a mean base quality of ~54 (Table 3.1). This indicates that the buffalo reference genome used performed well for the alignment of these sequences. The slight discrepancy between the duplicate removed and duplicate present alignments, with regards to the mean depth of coverage is due to the decrease in the amount of sequence present after duplicate removal. This led to the ~2x decrease in mean depth of coverage for the duplicate removed alignment (Table 3.2). For low coverage sequencing this depth of coverage is serviceable to perform further analysis on

the alignments [102]. This discrepancy in depth could have an effect on downstream processing and analysis when comparing the duplicates removed and present data. However, duplicate reads have become less of a problem with more modern sequencing platforms, such as the Illumina platform used in this study. Thus, the effect of not removing duplicates could be negligible for any downstream analysis. The most apparent effect removing duplicates had on the data was the aforementioned reduction in mean depth of coverage when comparing the population statistic results from the combined genotyping (CG) and combined genotyping duplicates removed (CGDR) pipelines.

4.2 Genotyping and variant calling

The individual genotyping (IG) pipeline produced the most called variants (~58 million SNPs) compared to the other two pipelines, which is likely due to the presence of duplicate reads and the genotyping/variant calling methods used for IG. However, without a dataset with duplicates removed and using individual genotype calculations, the true reason for this discrepancy in the number of variants is unclear and is outside the scope of this thesis. The same applies to any result with the same discrepancy between these pipelines. After filtering, the IG pipeline produced fewer SNPs that the other two pipelines, having only 9 million compared to the ~37. This was most likely due to the IG genotyping method calling variants which fall below the filter parameters. This is due to the lack of combined genotypic data, as would be derived from using a combined method of genotype calculation [84,88,93,103]. This includes variant statistics such as minor allele count and depth, which would be much higher were the variants to be called using multiple samples' data simultaneously. Filter parameters using frequency-based and depth-based metrics would also be affected in the same way, consequently reducing the number of

variants further. Thus, using the combined genotyping method would produce more variants which would otherwise be lost during this step of filtration. Consequently, variants which are more representative of the samples as a cohort would be lost unless all of the samples were to be genotyped at once. This is especially true if external population statistics are absent, such are those found in SNP databases (which is absent for the African buffalo). Therefore, to produce more accurate population statistics, the combined genotyping methods would be more appropriate [104,105]. Transitions are nucleotide substitutions from a purine to a different purine or a pyrimidine to a different pyrimidine (A to G or T to C, vice versa), while transversions are nucleotide substitutions from a purine to a pyrimidine (A/G to C/T) and vice versa. The optimal Ts/Tv for mammalian whole genome sequences is 2.10 [104]. This is because transitions are approximately twice as common as transversions in mammalian genomes (across the whole genome). The closer the ratio is to the optimal, the lower the chance of false positives variants are. The transition to transversion ratio (Ts/Tv) improved for all the data after filtering, which indicates an increase in overall variant quality. Even though the CG and CGDR pipelines had a Ts/Tv of 2.05, it is still in an acceptable range [80,88].

4.3 Inbreeding

Inbreeding is defined as the interbreeding of closely related individuals and can occur in a population due to several factors. These include behavioural factors, such as mating patterns and social structure [87,106–108]. Environmental factors can also influence inbreeding within a population by dictating its sizes and migration patterns (gene-flow). The inbreeding results indicated in general that there is a noticeable discrepancy between the results from PLINK and VCFtools (Figure 3.4). This is due to the two tools using

different methods of calculating the inbreeding coefficient (F) per individual. Even with the PLINK command set to the small sample mode, the sample size was still likely too small for the tool to function as intended with the method that PLINK uses (as described in the PLINK documentation) [89,90]. Curiously, when observed, the inbreeding plots from IG (Figure 3.4 a) resemble inversions of the results from the CG/CGDR plots. In other words, the level of inbreeding in IG for each individual relative to one another is approximately (based on visual observation of Figure 3.4) proportional to the results from CG and CGDR. This could mean that there is a possibility that the algorithm which calculates F for the IG dataset required an adjustment (such as cohort size and variant frequency) to account for the excess of heterozygosity produced by the IG pipeline, seeing as the inbreeding values were extremely negative (Figure 3.4a). However this is speculation, and does not constitute a reliable conclusion, based on the erroneous nature of negative inbreeding values [80,88]. As stated previously, the negative inbreeding results in Figure 3.4a were interpreted as erroneous.

As VCFtools inbreeding calculations do not have the limitation on sample size of PLINK, displaying few inbreeding coefficients below 0, and the ones that were indeed below 0 were negligibly so. Being mindful of this, all inbreeding results from the CG and CGDR form an approximate consensus of the levels of relatedness of each individual. The individuals from the AENP were the most inbred, having inbreeding coefficients within the range of second-degree relatedness (grandparent/grandchild level of relatedness) ~0.2-0.3. The individuals from HiP displayed moderate to high levels inbreeding with inbreeding values ranging from third to second degree relatedness (~0.1-0.2). One individual from KNP (B98_597) displayed high inbreeding (~0.15-0.20) compared to the

rest of the KNP individuals. This likely due to an isolated event of inbreeding between two related individuals since the rest of the KNP individuals showed low to no inbreeding. Similarly, all the MNP individuals showed low to no inbreeding.

The very low inbreeding values produced from the IG pipeline can be indicative of the aforementioned lack of appropriate genotypic information for the called SNPs. This caused there to be an excess of observed versus expected heterozygosity. This excess in heterozygosity could possibly also be attributed to an undiscovered flaw or error in the upstream portion of the IG pipeline. However, this is unlikely because the IG pipeline shares all the upstream data processing as the CG pipeline, as they both used the same BAM files to generate the genotype likelihoods and variant calling. Severe contamination at the library preparation or the sequencing level could also be the cause of severely negative inbreeding coefficients [80,88,89]. Whether or not there was contamination of the samples is inconclusive based on the discrepancy between the VCFtools and PLINK outputs. However, when considering the general quality of the sequences before and after processing and filtering, it is unlikely that severe contamination is present.

4.4 Population structure

There is clear population structure in the buffalo cohort used in this study, which can be observed in PLINK principal component analysis (PCA) plots Figure 3.5. Unlike the inbreeding statistics, the PCA effectively produced the same results for all three pipelines. The only discrepancy is that CG and CGDR produced tighter and more distantly separated clusters compared to IG. The buffalo cohort grouped into three distinct population clusters. The results reflected what was expected, based on the population

histories of the buffalo in those regions: The KNP and MNP individuals clustered together, as the MNP buffalo population was founded in 1999 to 2007 from disease free individuals from the KNP [109]. The separation of the other populations reflects the geographical separation of the parks and corresponds with the genetic separation of the populations observed by van Hooft et al (2000)[106]. This significant level of population structure would significantly confound any attempted genetic association unless properly adjusted for by making for three population clusters covariates or using only data from one of the clustered populations.

4.5 SNP panel

4.5.1 Variant statistics

After the population structure was determined, the CGDR pipeline's data was exclusively used for the subsequent analysis, as the pipeline produced theoretically could have produced highest quality variants, since duplicates were removed, along with the combined genotyping aspect of it[79]. This use of a single set of data also helped to focus further analysis, by reducing unnecessary complications, as this allowed the generation of a single full SNP panel. The variant statistics produced from CGDR, such as missingness, Hardy-Weinberg equilibrium (HWE), minor allele frequency (MAF) did not present any abnormal data [88]. Thus, standard variant filtering criteria with regards to those three statistics could be implemented for the creation of the SNP panel. Instead of using a standard MAF cut-off based on the MAF distribution (Figure 3.6 d) produced, the minimum MAF was based on the results from the statistical power test, which was performed using prior information (Chapter 2.8.4). This led to a minimum cut-off that was still considered to be standard (0.10). Statistics which could not be generated for this

cohort include sex-discrepancy for each individual. The statistic is based on the X-chromosome data, which was impossible to generate, because the reference genome used in this study did not have a chromosome level of assembly and is assembled into super-scaffolds. Also, excluding individuals during the SNP panel creation could have significant effects on the outcome due to the small size of the cohort.

4.5.2 Full SNP panel

The genes were selected for this panel based on previous associations with TB infection. Two of these genes, namely *IL1A* and *SLC13A1*, have been previously associated with bTB in African buffalo [63]. The remaining genes were all associated with TB in humans [56]. The genes which were associated in humans were selected due to this association and as they had orthologous counterparts in cattle. Some discrepancies in gene names between humans and cattle were noted, however, due to their orthologous nature this has no impact on the selection and function of these genes. The majority of these genes and their human counterparts can be seen in Figure 1.3. As can be seen in the figure, these genes represent a vast range of functions across several aspects of immunity [56,61,64,99]. Thus, this modest group of genes are suitably diverse and thereby representative for host genetics studies with regards to bTB. The number of variants relative to each of these genes was proportional to their size (Table 3.3). Unfortunately, all 3698 SNPs could not be annotated fully as it would have been too laborious and time consuming and would outweigh its relative use for this study.

4.5.3 Targeted genes

The genes that were selected to be included in the panel were three immune related genes, which have previously been associated with disease susceptibility in bovids. *IL1a*

and *SLC7A13* where both associated with bTB susceptibility in African buffalo in the study done by Le Roex et al. (2013). *IL1a* encodes for the interleukin-1-alpha protein (IL-1a) which is primarily produced by macrophages. It is involved in the inflammatory response and involved in several immune pathways, primarily the IL-1 signalling pathway. It serves to increase the proliferation of the precursor cells to T-cells in the thymus (thymocytes). The *SLC7A13* gene encodes for the solute carrier family 7-member 13 protein, which is a transmembrane solute carrier protein. The gene and its protein are relatively poorly characterized, however the closely related solute carrier family 11 member 1 (SLC11A1) was previously associated with TB susceptibility in humans as well as bTB susceptibility in cattle [65,110]. It is thus likely that the *SLC7A13* could have a similar association to bTB susceptibly in buffalo. The last gene, *BoLA-DRB3,* is the bovine equivalent of major histocompatibility complex (MHC) class II genes in other mammals such as humans. These genes encode for several products which are involved in adaptive immunity, primarily cell surface proteins involved in antigen-presentation. This is crucial for detecting, tagging and clearing pathogens. The class II genes are primarily expressed in CD4 T cells, which, as explained in Chapter 1.8, play an important role in TB immunity and susceptibility.

4.5.4 Final SNP panel

After using the filtration parameters, 143 high quality SNPs located in the selected gene regions were obtained. These SNPs would serve as a panel of markers for association tests for bTB susceptibility in the African buffalo. To further identify possible candidate SNPs for association, the functional consequence of the SNPs was investigated. The effects of the SNPs were predicted in the cow genome as an analogue to the buffalo due

to the lack of available buffalo information available at the time. The majority (108) of these SNPs were found to be wildtype in the cow (reference genome ARS-UCD1.2, GCF_002263795.1), according to Ensembl's Variant Effect Predictor (VEP) tool. VEP does not present any results for variants which are wildtype in its output, but there were 36 SNPs with variant information. Of these 36 SNPs, 33 were novel and 3 were previously observed, only one of which was a nonsynonymous substitution in a protein coding region of *BoLA-DRB3*. This variant displayed a low level of predicted pathogenicity (SIFT score: 0.12), indicating that the effect of the SNP is likely tolerated. The SNP does however occur at a highly conserved locus (Figure 3.7 b). This could mean that it has an effect which the predictive pathogenicity does not adequately represent, which is always a possibility due to the relative inaccuracy of predictive statistics such as these. The SNP causes a isoleucine to valine amino acid substitution in the immunoglobulin C-Type (IGc1) domain of the protein product (Figure 4.1)[111,112]. This domain is a C-1 type domain which is involved almost entirely in immune functions and is primarily found in molecules produced by the MHC genes and other T cell surface receptors [111,112]. Disruptions in this domain can thus result in the reduction or absence of function of the proteins associated with it and by extension can lead to reduced T cell immune functions. The amino acid substitution is however likely tolerated due to the change not causing any major structural changes, as both isoleucine and valine are non-polar amino acids with similar side change functions. SMART determined that the variant does not cause any change in the domain structure or function. Thus, it can be concluded that there is only a small likelihood that this SNP could have a significant deleterious effect on the gene's function.

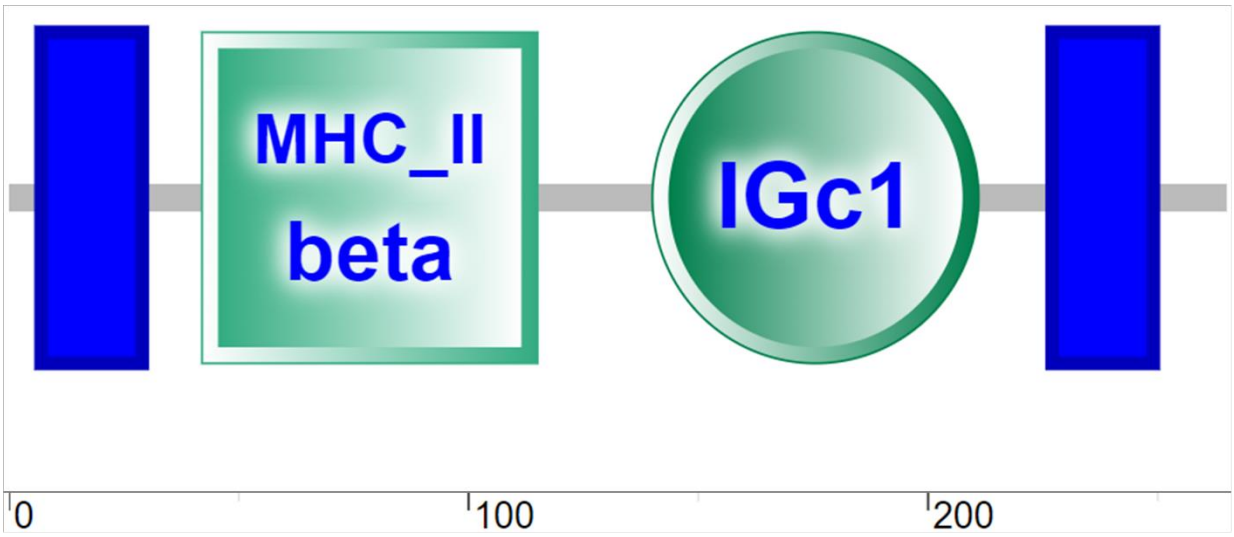


Figure 4.1: The protein domains of the BoLA-DRB3 protein as represented by SMART (SMART ID: Q3SZG0_BOVIN). The square ladled MHC_II beta represents the class II histocompatibility antigen, beta domain. The circle ladled IGc1 represents the immunoglobulin C-Type (IGc1) domain. The two blue rectangles represent trans membrane regions. The bar at the bottom represents the amino acid positions.

Regardless of all the VEP results and possible consequences the SNPs might have, all of them were included in the final panel. This is because even if the SNPs have no significant effect themselves, they could still serve as markers for true causative variants or loci through linkage. Even the SNPs that were found to be wildtype in the cow genome were kept, because these variants remain non-wildtype in the buffalo, which could still carry significant effects yet to be discovered. It should be noted that none of the variants found in this panel were found in the study by Le Roex et al. (2013)[63].

4.6 Targeted association test

The SNP panel was then tested to see how it would perform in a basic association analysis. This was done using the available data by testing the cohort of 40 buffalo samples for association of the SNPs with bTB exposure. The association was severely statically underpowered, due to the small sample size and skewed cases to controls ratio

(30:10). Thus, none of the results would be considered truly significant, nevertheless the results could still point towards a potential trend that could serve as the basis for further, more robust analysis. The cases were all considered bTB exposed as they were from the two bTB endemic populations (KNP and HiP), while the controls were the individuals from the bTB free populations (AENP and MNP). The test could not account for all the confounding factors which could influence the results, due to the lack of the appropriate data to account for them. This data includes true accurate molecular genotypic data for the individuals. Individuals were also not removed based on the inbreeding values or other factors due to the already small sample size which would reduce the validity of the test even further. The association test using a Cochran-Armitage trend test resulted in 10 significantly associated ($p<0.05$) SNPs with bTB exposure (Table 3.4). Most of these SNPs were in *SLC7A13*. Which could possibly indicate that the gene is likely to be associated with bTB susceptibility in African buffalo. The two other significantly associated SNPs were located in *BoLA-DRB3*. None of the SNPs which showed significantly association had VEP consequence information as they were all wildtype in the cow, they were also all located in intronic gene regions. Thus, little information could be gathered on the possible reasons for these associations. The prevailing reason can always be, as previously mentioned, that these SNPs act as markers for true causative variants or loci.

4.7 Limitations

The primary limitation of this study was the lack of reference genome information that could be used to analyse the 40 Cape buffalo samples. The African buffalo reference genome produced by Glanzmann et al. (2016)[7] is a high-quality reference, however the

relatively large amount of scaffold and contigs in the assembly lead to the use of the super scaffold stitched version of the reference. This rendered the coordinate-based annotations for the reference genome unusable, which lead to the reliance of the cow reference genome to annotate variants that were found in the buffalo. Using the cow reference for gene locations and annotations could produce far less accurate results than if the appropriate buffalo annotations had been used. In general, studying poorly characterized organisms will have similar limitations, where the lack of available data hampers the processing and accurate interpretation of certain genetic elements. The limitation extended to the limited number of genes selected for the panel relative to amount of the whole genome data available. For the association test subset panel, the conversion of the buffalo loci coordinates to the cow coordinates had to be done manually using the BLAST outputs which were generated for the genes. This process was extremely inefficient, due to the lack of any form of computational automation or software tool that could be used for this. This limited the number of genes that were included in the test panel.

The original intent was to find/implement the most efficient variant calling pipeline for obtaining and analysing the SNPs form the 40 samples in the study cohort. To this end, two tool sets were investigated and used, namely the Genome Analysis Toolkit (GATK, version 4.0.6.0) with HaplotypeCaller [81,113], which is standard practice, and SAMtools/BCFtools. However, after extensive troubleshooting of the tools for use in this study, which involved their use in multiple configurations, it was ultimately decided to make use of BCFtools in this study, as it was shown to be the most efficient tool. The quality control measures implemented in this study was also implemented to try and

mitigate the impact this would have on the results. Consequently, a major limitation of this study is that HaplotypeCaller was not used for genotyping/variant calling. The use of the tool is preferable due to the high confidence its base calling algorithm affords in each variant it calls, especially around insertion/deletion (INDEL) sites. This is due to its use of a haplotype-based calling algorithm, which incorporates realignment of the input alignment and iterative base calling [113]. It has also been shown in several studies to produce higher confidence variant calls [114–116], and if this project would be repeated, HaplotypeCaller should be implemented.

Another main limitation of this study is the small sample size. The use of such few samples limits all population-based analysis, such as heterozygosity or inbreeding/ relatedness. For a non-model species such as the African buffalo, this can also mean that there is a lot less variant frequency information available to distinguish rare and common variants in the population. It also significantly reduces the power of any association-based analysis. There are also several variant and individual based statistics that cannot be effectively implemented for SNP quality control filtration to make a SNP panel. This includes statistics such as the one based on heterozygosity, genotypic data missingness, heritability and HWE.

The use of low coverage sequencing (even though it was of high quality) versus high-coverage sequencing impaired the statistical analyses reliant on depth of coverage, such as inbreeding [87,109,117]. An additional limitation was that there were no individuals confirmed to be bTB infected included in this study. Only a subset of cohort was considered to be exposed to the disease at the time of recruitment. Thus, with regards to the association test, no distinction could be made between true susceptibility versus

resistance to bTB. All these limitations acted in a compounding manner to affect the final association test. The test lacked statistical power to produce any significant association between the SNPs and bTB susceptibility. However, this was not a primary aim for this thesis, as future studies with a larger cohort of buffalo and selected cases and controls would have had to been used for a statistically viable association test.

4.8 Future applications and research

The SNP panel produced is of high quality and containing marker loci for a potential follow up study aimed at the genetic association of bTB susceptibility in African buffalo. However, any future study would have to consider the key limitations described in this thesis. The use of a new and refined reference genome assembled to the chromosome level as well as a suitably large cohort would ensure that the outcome is improved over what is presented here. Not having to rely on other reference species, such as the cow, will increase the accuracy and confidence which one can have in the results significantly. Assembling more variant information for the African buffalo would also increase the general effectiveness of any genetic and population-based study performed in this context. In the case where the SNPs produced in this study is used, a good starting point would be to obtain the physical genotypic data for all the individuals in the new study population for the SNP loci present in the panel. This data will not only confirm the true genotypes of the individuals for those loci, but also provide the data required to perform an association test for bTB susceptibility. The use of new high-coverage whole genome sequencing would allow more SNPs and genes to be added to the panel. Further, linkage studies can be performed on such data to elucidate any potentially linked markers or causative variants associated with bTB susceptibility. For a truly effective association test

however, more thorough and accurate population statistics should be obtained, using the same methodology as presented here, but with the larger cohort. This would be to properly account for population structure and relatedness in the study population to limit these factors' confounding effects on the association.

4.9 Conclusions

This study produced a panel of 3698 high-quality SNPs which can serve as a basis for targeted or marker assisted studies in African buffalo specifically for bTB susceptibility. The low-coverage whole genome sequences used in this study were of high quality and with the limited amount of data, still produce serviceable population statistics. However, any statistics produced here are subject to the limitations presented in this thesis and should thus be further investigated before a clearer and more accurate conclusion can be made. These limitations pertain mainly to the sample size and lack of available variant annotation data for the African buffalo. Nevertheless, the results of this study can serve as a basis for future studies, allowing the use of the SNP panel to identify possible genetic factors which contribute to bTB susceptibility in African buffalo. This can in turn aid in the development of bTB control techniques through either diagnostic or treatment avenues. This can by extension improve our understanding of TB immunity in all mammalian hosts of the disease, including humans.

References

1. Fitzgerald SD, Kaneene JB. Wildlife Reservoirs of Bovine Tuberculosis Worldwide: Hosts, Pathology, Surveillance, and Control. *Vet Pathol*. 2013;50(3):488-499. doi:10.1177/0300985812467472

2. Meunier N V., Sebulime P, White RG, Kock R. Wildlife-livestock interactions and risk areas for cross-species spread of bovine tuberculosis. *Onderstepoort J Vet Res*. 2017;84(1):1-10. doi:10.4102/ojvr.v84i1.1221

3. Finlay EK, Berry DP, Wickham B, Gormley EP, Bradley DG. A genome wide association scan of bovine tuberculosis susceptibility in Holstein-Friesian dairy cattle. *PLoS One*. 2012;7(2). doi:10.1371/journal.pone.0030545

4. Robinson PA. Framing bovine tuberculosis: a 'political ecology of health' approach to circulation of knowledge(s) about animal disease control. *Geogr J*. 2017;183(3):285-294. doi:10.1111/geoj.12217

5. Hoffman LC, van As JS, Gouws PA, Govender D. Carcass Yields of African Savanna Buffalo (Syncerus caffer caffer). *African J Wildl Res*. 2020;50(1):69. doi:10.3957/056.050.0069

6. Lepori AA, Josling GC, Neser FWC, Lubout PC, van Wyk JB. Multi-trait genetic evaluation for horn traits of economic importance in the Cape buffalo (Syncerus caffer caffer). *South African J Anim Sci*. 2019;49(2):364-370. doi:10.4314/sajas.v49i2.15

7. Glanzmann B, Möller M, le Roex N, Tromp G, Hoal EG, van Helden PD. The

complete genome sequence of the African buffalo (Syncerus caffer). *BMC Genomics*. 2016;17(1):1-7. doi:10.1186/s12864-016-3364-0

8. Winnie JA, Cross P, Getz W, Winnie JA, Cross P, Getz W. Habitat Quality and Heterogeneity Influence Distribution and Behavior in African Buffalo Published by : Wiley on behalf of the Ecological Society of America Stable URL : http://www.jstor.org/stable/27651689 REFERENCES Linked references are available on JS. 2017;89(5):1457-1468.

9. Smitz N, Cornélis D, Chardonnet P, et al. Genetic structure of fragmented southern populations of African Cape buffalo (Syncerus caffer caffer) Hirohisa Kishino. *BMC Evol Biol*. 2014;14(1):1-19. doi:10.1186/s12862-014-0203-2

10. IUCN SSC Antelope Specialist Group. Syncerus caffer. *IUCN Red List Threat Species 2019 eT21251A50195031*. Published online 2019. http://dx.doi.org/10.2305/IUCN.UK.2019-1.RLTS.T21251A50195031.en

11. le Roex N, Noyes H, Brass A, et al. Novel SNP Discovery in African Buffalo, Syncerus caffer, Using High-Throughput Sequencing. *PLoS One*. 2012;7(11):4-9. doi:10.1371/journal.pone.0048792

12. Grange JM, Yates MD. Guidelines for speciation within the Mycobacterium tuberculosis complex. *World Heal Organ*. 1996;2:1-23. http://whqlibdoc.who.int/hq/1996/WHO_EMC_ZOO_96.4.pdf

13. Zhou H, Sinsheimer JS, Bates DM, et al. OpenMendel: a cooperative programming project for statistical genetics. *Hum Genet*. 2020;139(1):61-71. doi:10.1007/s00439-019-02001-z

14. VerCauteren KC, Lavelle MJ, Campa H. Persistent spillback of bovine tuberculosis from white-tailed deer to cattle in Michigan, USA: Status, Strategies, and Needs. *Front Vet Sci*. 2018;5(NOV):1-13. doi:10.3389/fvets.2018.00301

15. Srinivasan S, Easterling L, Rimal B, et al. Prevalence of Bovine Tuberculosis in India: A systematic review and meta-analysis. *Transbound Emerg Dis*. 2018;65(6):1627-1640. doi:10.1111/tbed.12915

16. Renwick AR, White PCL, Bengis RG. Bovine tuberculosis in southern African wildlife: A multi-species host-pathogen system. *Epidemiol Infect*. 2007;135(4):529-540. doi:10.1017/S0950268806007205

17. Miller MA, Buss P, Parsons SDC, et al. Conservation of white rhinoceroses threatened by bovine tuberculosis, South Africa, 2016–2017. *Emerg Infect Dis*. 2018;24(12):2373-2375. doi:10.3201/eid2412.180293

18. Olea-Popelka F, Muwonge A, Perera A, et al. Zoonotic tuberculosis in human beings caused by Mycobacterium bovis—a call for action. *Lancet Infect Dis*. 2017;17(1):e21-e25. doi:10.1016/S1473-3099(16)30139-6

19. Domingo M, Vidal E, Marco A. Pathology of bovine tuberculosis. *Res Vet Sci*. 2014;97(S):S20-S29. doi:10.1016/j.rvsc.2014.03.017

20. Russell DG, Barry CE, Flynn JL. Tuberculosis : What We Don ' t Know. *Science (80-)*. 2010;328(5980):852-856. http://science.sciencemag.org/content/328/5980/852.full

21. de Martino M, Lodi L, Galli L, Chiappini E. Immune Response to Mycobacterium

tuberculosis: A Narrative Review. *Front Pediatr*. 2019;7(August):1-8. doi:10.3389/fped.2019.00350

22. Bernitz N, Kerr TJ, Goosen WJ, et al. Impact of Mycobacterium bovis-induced pathology on interpretation of QuantiFERON®-TB Gold assay results in African buffaloes (Syncerus caffer). *Vet Immunol Immunopathol*. 2019;217(June):109923. doi:10.1016/j.vetimm.2019.109923

23. Dejene SW, Heitkönig IMA, Prins HHT, et al. Risk factors for bovine tuberculosis (bTB) in cattle in Ethiopia. *PLoS One*. 2016;11(7):1-17. doi:10.1371/journal.pone.0159083

24. Ahmady EB. Some Aspects about the Bovine Tuberculosis. *J Zoo Biol*. 2018;1(1):29-41. doi:10.33687/zoobiol.001.01.1004

25. Ghebremariam MK, Michel AL, Nielen M, Vernooij JCM, Rutten VPMG. Farm-level risk factors associated with bovine tuberculosis in the dairy sector in Eritrea. *Transbound Emerg Dis*. 2018;65(1):105-113. doi:10.1111/tbed.12622

26. Ayele WY, Neill SD, Zinsstag J, Weiss MG, Pavlik I. Bovine tuberculosis: An old disease but a new threat to Africa. *Int J Tuberc Lung Dis*. 2004;8(8):924-937.

27. Michel AL, Bengis RG. The African buffalo: A villain for inter-species spread of infectious diseases in southern Africa. *Onderstepoort J Vet Res*. 2012;79(2). doi:10.4102/ojvr.v79i2.453

28. Dawson KL, Stevenson MA, Sinclair JA, Bosson MA. Recurrent bovine tuberculosis in New Zealand cattle and deer herds, 2006-2010. *Epidemiol Infect*.

2014;142(10):2065-2074. doi:10.1017/S0950268814000910

29. Rivière J, Le Strat Y, Hendrikx P, Dufour B. Cost-effectiveness evaluation of bovine tuberculosis surveillance in wildlife in France (Sylvatub system) using scenario trees. *PLoS One*. 2017;12(8). doi:10.1371/journal.pone.0183126

30. Bermingham ML, More SJ, Good M, et al. Genetics of tuberculosis in Irish Holstein-Friesian dairy herds. *J Dairy Sci*. 2009;92(7):3447-3456. doi:10.3168/jds.2008-1848

31. Payne A, Boschiroli ML, Gueneau E, et al. Bovine tuberculosis in “Eurasian” badgers (Meles meles) in France. *Eur J Wildl Res*. 2013;59(3):331-339. doi:10.1007/s10344-012-0678-3

32. Laubscher L, Hoffman L. An overview of disease-free buffalo breeding projects with reference to the different systems used in South Africa. *Sustainability*. 2012;4(11):3124-3140. doi:10.3390/su4113124

33. World Health Organization. *Global Tuberculosis Report 2018. World Health Organization. Http://Www.Who.Int/Iris/Handle/10665/274453.*; 2018.

34. Mamo G, Abebe F, Worku Y, Hussein N. Bovine tuberculosis and its associated risk factors in pastoral and agro-pastoral cattle herds of Afar Region , Northeast Ethiopia. *J Vet Med Anim Heal*. 2013;5(June):171-179. doi:10.5897/JVMAH2013.0204

35. Musoke J, Hlokwe T, Marcotty T, Plessis BJA, Michel AL. Spillover of Mycobacterium bovis from Wildlife to Livestock, South Africa. *Emerg Infect Dis*.

2015;21(3):448-451.

36. Bernitz N, Kerr TJ, Goosen WJ, et al. Review of Diagnostic Tests for Detection of Mycobacterium bovis Infection in South African Wildlife. *Front Vet Sci*. 2021;8(January):1-11. doi:10.3389/fvets.2021.588697

37. Maas M, Michel AL, Rutten VPMG. Facts and dilemmas in diagnosis of tuberculosis in wildlife. *Comp Immunol Microbiol Infect Dis*. 2013;36(3):269-285. doi:10.1016/j.cimid.2012.10.010

38. Palmer M V. Mycobacterium bovis: Characteristics of wildlife reservoir hosts. *Transbound Emerg Dis*. 2013;60(SUPPL1):1-13. doi:10.1111/tbed.12115

39. White PCL, Böhm M, Marion G, Hutchings MR. Control of bovine tuberculosis in British livestock: there is no “silver bullet.” *Trends Microbiol*. 2008;16(9):420-427. doi:10.1016/j.tim.2008.06.005

40. Buddle BM, Vordermeier HM, Chambers MA, de Klerk-Lorist LM. Efficacy and safety of BCG vaccine for control of tuberculosis in domestic livestock and wildlife. *Front Vet Sci*. 2018;5(OCT):1-17. doi:10.3389/fvets.2018.00259

41. Ramsey DSL, Efford MG. Management of bovine tuberculosis in brushtail possums in New Zealand: Predictions from a spatially explicit, individual-based model. *J Appl Ecol*. 2010;47(4):911-919. doi:10.1111/j.1365-2664.2010.01839.x

42. Riordan P, Delahay RJ, Cheeseman C, Johnson PJ, Macdonald DW. Culling-induced changes in badger (Meles meles) behaviour, social organisation and the epidemiology of bovine tuberculosis. *PLoS One*. 2011;6(12).

doi:10.1371/journal.pone.0028904

43. Palmer M V., Thacker TC, Waters WR, Robbe-Austerman S. Oral vaccination of white-tailed deer (Odocoileus virginianus) with Mycobacterium bovis Bacillus Calmette-Guerin (BCG). *PLoS One*. 2014;9(5):1-6. doi:10.1371/journal.pone.0097031

44. Goosen WJ, Kerr TJ, Kleynhans L, et al. The VetMAX™ M. tuberculosis complex PCR kit detects MTBC DNA in antemortem and postmortem samples from white rhinoceros (Ceratotherium simum), African elephants (Loxodonta africana) and African buffaloes (Syncerus caffer). *BMC Vet Res*. 2020;16(1):16-21. doi:10.1186/s12917-020-02438-9

45. Bernitz N, Clarke C, Roos EO, et al. Detection of Mycobacterium bovis infection in African buffaloes (Syncerus caffer) using QuantiFERON®-TB Gold (QFT) tubes and the Qiagen cattletype® IFN-gamma ELISA. *Vet Immunol Immunopathol*. 2018;196(December 2017):48-52. doi:10.1016/j.vetimm.2017.12.010

46. Roos EO, Olea-Popelka F, Buss P, et al. Seroprevalence of Mycobacterium bovis infection in warthogs (Phacochoerus africanus) in bovine tuberculosis-endemic regions of South Africa. *Transbound Emerg Dis*. 2018;65(5):1182-1189. doi:10.1111/tbed.12856

47. Hlokwe TM, van Helden P, Michel AL. Evidence of increasing intra and inter-species transmission of Mycobacterium bovis in South Africa: Are we losing the battle? *Prev Vet Med*. 2014;115(1-2):10-17. doi:10.1016/j.prevetmed.2014.03.011

48. Goosen WJ, Kerr TJ, Kleynhans L, et al. The Xpert MTB/RIF Ultra assay detects

Mycobacterium tuberculosis complex DNA in white rhinoceros (Ceratotherium simum) and African elephants (Loxodonta africana). *Sci Rep*. 2020;10(1):1-7. doi:10.1038/s41598-020-71568-9

49. Kao SYZ, VanderWaal K, Enns EA, et al. Modeling cost-effectiveness of risk-based bovine tuberculosis surveillance in Minnesota. *Prev Vet Med*. 2018;159(April):1-11. doi:10.1016/j.prevetmed.2018.08.011

50. de la Rua-Domenech R, Goodchild AT, Vordermeier HM, Hewinson RG, Christiansen KH, Clifton-Hadley RS. Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, γ-interferon assay and other ancillary diagnostic techniques. *Res Vet Sci*. 2006;81(2):190-210. doi:10.1016/j.rvsc.2005.11.005

51. Allen AR, Minozzi G, Glass EJ, et al. Bovine tuberculosis: The genetic basis of host susceptibility. *Proc R Soc B Biol Sci*. 2010;277(1695):2737-2745. doi:10.1098/rspb.2010.0830

52. Bernitz N, Kerr TJ, Goosen WJ, et al. Parallel measurement of IFN-γ and IP-10 in QuantiFERON®-TB Gold (QFT)plasma improves the detection of Mycobacterium bovis infection in African buffaloes (Syncerus caffer). *Prev Vet Med*. 2019;169(May):104700. doi:10.1016/j.prevetmed.2019.104700

53. Raphaka K, Sánchez-Molano E, Tsairidou S, et al. Impact of genetic selection for increased cattle resistance to bovine tuberculosis on disease transmission dynamics. *Front Vet Sci*. 2018;5(OCT):1-14. doi:10.3389/fvets.2018.00237

54. Raphaka K, Matika O, Sánchez-Molano E, et al. Genomic regions underlying

susceptibility to bovine tuberculosis in Holstein-Friesian cattle. *BMC Genet*. 2017;18(1):27. doi:10.1186/s12863-017-0493-7

55. Wilkinson S, Bishop SC, Allen AR, et al. Fine-mapping host genetic variation underlying outcomes to Mycobacterium bovis infection in dairy cows. *BMC Genomics*. 2017;18(1):1-13. doi:10.1186/s12864-017-3836-x

56. Möller M, Hoal EG. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis*. 2010;90(2):71-83. doi:10.1016/j.tube.2010.02.002

57. Amos W, Brooks-Pollock E, Blackwell R, Driscoll E, Nelson-Flower M, Conlan AJK. Genetic Predisposition to Pass the Standard SICCT Test for Bovine Tuberculosis in British Cattle. *PLoS One*. 2013;8(3). doi:10.1371/journal.pone.0058245

58. Muller B, Durr S, Hattendorf J, et al. Zoonotic Mycobacterium bovis-induced tuberculosis in humans. *Emerg Infect Dis* . 2013;19(6):899-908.

59. Bhuju S, Aranday-Cortes E, Villarreal-Ramos B, Xing Z, Singh M, Vordermeier HM. Global Gene Transcriptome Analysis in Vaccinated Cattle Revealed a Dominant Role of IL-22 for Protection against Bovine Tuberculosis. *PLoS Pathog*. 2012;8(12). doi:10.1371/journal.ppat.1003077

60. Phillips CJC, Foster CRW, Morris PA, Teverson R. Genetic and management factors that influence the susceptibility of cattle to Mycobacterium bovis infection . *Anim Heal Res Rev*. 2002;3(1):3-13. doi:10.1079/ahrr200236

61. Ghamari E, Farnia P, Saif S, et al. Susceptibility to pulmonary tuberculosis: host genetic deficiency in tumor necrosis factor alpha (TNF-α) gene and tumor necrosis factor receptor 2 (TNFR2). *Int J Mycobacteriology*. 2016;5:S136-S137. doi:10.1016/j.ijmyco.2016.09.038

62. Queirós J, Alves PC, Vicente J, Gortázar C, De La Fuente J. Genome-wide associations identify novel candidate loci associated with genetic susceptibility to tuberculosis in wild boar. *Sci Rep*. 2018;8(1):1-12. doi:10.1038/s41598-018-20158-x

63. le Roex N, Koets AP, van Helden PD, Hoal EG. Gene Polymorphisms in African Buffalo Associated with Susceptibility to Bovine Tuberculosis Infection. *PLoS One*. 2013;8(5):1-6. doi:10.1371/journal.pone.0064494

64. Eirin M, Carignano H, Shimizu E, et al. BoLA-DRB3 exon2 polymorphisms among tuberculous cattle: Nucleotide and functional variability and their association with bovine tuberculosis pathology. *Res Vet Sci*. 2020;130(May 2019):118-125. doi:10.1016/j.rvsc.2020.03.001

65. Holder A, Garty R, Elder C, et al. Analysis of Genetic Variation in the Bovine SLC11A1 Gene, Its Influence on the Expression of NRAMP1 and Potential Association With Resistance to Bovine Tuberculosis. *Front Microbiol*. 2020;11(June):1-9. doi:10.3389/fmicb.2020.01420

66. van Hooft P, Dougherty ER, Getz WM, Greyling BJ, Zwaan BJ, Bastos ADS. Genetic responsiveness of African buffalo to environmental stressors: A role for epigenetics in balancing autosomal and sex chromosome interactions? *PLoS*

One. 2018;13(2):1-24. doi:10.1371/journal.pone.0191481

67. Smitz N, Berthouly C, Cornélis D, et al. Pan-African Genetic Structure in the African Buffalo (Syncerus caffer): Investigating Intraspecific Divergence. *PLoS One*. 2013;8(2). doi:10.1371/journal.pone.0056235

68. Leese F, Brand P, Rozenberg A, et al. Exploring Pandora's Box: Potential and Pitfalls of Low Coverage Genome Surveys for Evolutionary Biology. *PLoS One*. 2012;7(11). doi:10.1371/journal.pone.0049202

69. Stevenson KR, Coolon JD, Wittkopp PJ. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*. 2013;14(1). doi:10.1186/1471-2164-14-536

70. Waples RK, Albrechtsen A, Moltke I. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol Ecol*. 2019;28(1):35-48. doi:10.1111/mec.14954

71. Li Z, Meng M, Li S, Deng B. The transcriptome analysis of Protaetia brevitarsis Lewis larvae. *PLoS One*. 2019;14(3). doi:10.1371/journal.pone.0214001

72. Haj A. ScaffoldStitcher. Published online 2016. doi:10.5281/zenodo.3475473

73. Andrews S. FastQC: A quality control tool for high throughput sequence data. Published online 2010.

74. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354

75. Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*. 2012;28(14):1838-1844. doi:10.1093/bioinformatics/bts280

76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

77. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

78. Picard toolkit. *Broad Institute, GitHub Repos*. Published online 2019. http://broadinstitute.github.io/picard/

79. Ebbert MTW, Wadsworth ME, Staley LA, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. 2016;17(July). doi:10.1186/s12859-016-1097-3

80. Li H, Wren J. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 2014;30(20):2843-2851. doi:10.1093/bioinformatics/btu356

81. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110

82. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools.

Bioinformatics. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

83. Kishikawa T, Momozawa Y, Ozeki T, et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep*. 2019;9(1):1-10. doi:10.1038/s41598-018-38346-0

84. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5(9):1564-1573. doi:10.1038/nprot.2010.116

85. Joiret M, Mahachie John JM, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Min*. 2019;12(1):1-23. doi:10.1186/s13040-019-0199-7

86. Medina-Gomez C, Felix JF, Estrada K, et al. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur J Epidemiol*. 2015;30(4):317-330. doi:10.1007/s10654-015-9998-4

87. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. *Am J Hum Genet*. 2004;74(1):106-120. doi:10.1086/381000

88. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27(2):1-10. doi:10.1002/mpr.1608

89. Purcell S, Chang C. PLINK. www.cog-genomics.org/plink/1.9/

90. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7. doi:10.1186/s13742-015-0047-8

91. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2020. https://www.r-project.org/

92. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686

93. Lunetta KL. Genetic association studies. *Circulation*. 2008;118(1):96-101. doi:10.1161/CIRCULATIONAHA.107.700401

94. Hong EP, Park JW. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform*. 2012;10(2):117. doi:10.5808/gi.2012.10.2.117

95. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. *Nat Protoc*. 2011;6(2):121-133. doi:10.1038/nprot.2010.182

96. Johnson JL. This Genetic Association Study (GAS) Power Calculator. Published 2017. http://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html

97. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

98. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421

99. MAILLARD JC, MARTINEZ D, BENSAID A. An Amino Acid Sequence Coded by the Exon 2 of the BoLA DRB3 Gene Associated with a BoLA Class I Specificity Constitutes a Likely Genetic Marker of Resistance to Dermatophilosis in Brahman Zebu Cattle of Martinique (FWI)a. *Ann N Y Acad Sci*. 1996;791(1):185-197. doi:10.1111/j.1749-6632.1996.tb53525.x

100. Nassiry MR, Shahroodi FE, Mosafer J, et al. Analysis and frequency of bovine lymphocyte antigen (BoLA-DRB3) alleles in Iranian Holstein cattle. *Genetika*. 2005;41(6):817-822. http://www.ncbi.nlm.nih.gov/pubmed/16080607

101. Xu A, van Eijk MJ, Park C, Lewin HA. Polymorphism in BoLA-DRB3 exon 2 correlates with resistance to persistent lymphocytosis caused by bovine leukemia virus. *J Immunol*. 1993;151(12):6977-6985. http://www.ncbi.nlm.nih.gov/pubmed/8258704

102. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera A V. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *bioRxiv*. Published online 2019:1-12. doi:10.1101/716977

103. Sun Y V., Kardia SLR. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. *Eur J Hum Genet*. 2008;16(4):487-495. doi:10.1038/sj.ejhg.5201988

104. Depristo MA, Banks E, Poplin R, et al. A framework for variation discovery and

genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-501. doi:10.1038/ng.806

105. Smitz N, Van Hooft P, Heller R, et al. Genome-wide single nucleotide polymorphism (SNP) identification and characterization in a non-model organism, the African buffalo (Syncerus caffer), using next generation sequencing. *Mamm Biol*. 2016;81(6):595-603. doi:10.1016/j.mambio.2016.07.047

106. Van Hooft WF, Groen AF, Prins HHT. Microsatellite analysis of genetic diversity in African buffalo (Syncerus caffer) populations troughout Africa. *Mol Ecol*. 2000;9(12):2017-2025. doi:10.1046/j.1365-294X.2000.01101.x

107. Su SY, Kasberger J, Baranzini S, et al. Detection of identity by descent using next-generation whole genome sequencing data. *BMC Bioinformatics*. 2012;13(1). doi:10.1186/1471-2105-13-121

108. Benjelloun B, Boyer F, Streeter I, et al. An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. *Mol Ecol Resour*. 2019;19(6):1497-1515. doi:10.1111/1755-0998.13070

109. Winterbach HEK. The status and distribution of Cape buffalo Syncerus caffer caffer in southern Africa. *African J Wildl Res*. 1998;28(3):82-88.

110. Kim H-M, Jeon S, Chung O, et al. Comparative analysis of seven short-reads sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. Published online 2020. doi:10.1101/2020.03.22.002840

111. Teichmann SA, Chothia C. Immunoglobulin superfamily proteins in Caenorhabditis elegans 1 1Edited by G. von Heijne. *J Mol Biol*. 2000;296(5):1367-1383. doi:10.1006/jmbi.1999.3497

112. Radosevich M, Ono SJ. Novel Mechanisms of Class II Major Histocompatibility Complex Gene Regulation. *Immunol Res*. 2003;27(1):85-106. doi:10.1385/IR:27:1:85

113. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. Published online 2017:1-22. doi:10.1101/201178

114. Disratthakit A, Toyo-oka L, Thawong P, et al. An optimized genomic VCF workflow for precise identification of Mycobacterium tuberculosis cluster from cross-platform whole genome sequencing data. *Infect Genet Evol*. 2020;79(August 2019):104152. doi:10.1016/j.meegid.2019.104152

115. Pirooznia M, Kramer M, Parla J, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics*. 2014;8(1):14. doi:10.1186/1479-7364-8-14

116. Hofmann AL, Behr J, Singer J, et al. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics*. 2017;18(1):1-15. doi:10.1186/s12859-016-1417-7

117. Bi C, Lu N, Han T, et al. Whole-Genome Resequencing of Twenty Branchiostoma belcheri Individuals Provides a Brand-New Variant Dataset for Branchiostoma. *Biomed Res Int*. 2020;2020:7-9. doi:10.1155/2020/3697342