Towards sustainable agriculture:

Utilising genetics to sustainably improve the efficiency and longevity of cattle agriculture on both large-scale and localised level farming, in view of climate change effects.

A thesis to the University of Kent for the degree of:
MSc by Research in Genetics, School of Biosciences.

Carla Hazel Finn, 2022

Total word count: 15815, excl.references: 13687
Total pages: 55, excl.references: 49

Abstract (full abstract on page 9).

Livestock products have high densities of critical nutrients. As the world population increases, so too does the demand for animal products (such as milk), whilst climate change induced stressors are expected to intensify competition for resources, indicating that livestock systems must increase climatic resilience, alongside productivity and efficiency. The genetic diversity of livestock breeds is shaped by evolutionary forces such as genetic drift, migration, selection and geographical separation. Modern livestock genetics have also been influenced by human-mediated selection. As a result, many highly specialized breeds are adapted to localised environments as well as having evolved to meet a variety of human needs. Both processes have left traces in the genome of domestic livestock species as genome-wide variants such as short-nucleotide polymorphisms, 'SNPs'. The growing availability of computationally efficient genomic tools means that selection signatures can be readily analysed to assess the genetic diversity and population structure in cattle breeds and among cattle populations. This research utilised genetics in consideration of the growing need to safeguard livestock populations, by considering the need to increase efficiency as well as climate-change induced resilience, with a focus on cattle. We investigated and analysed the population structure and diversity of South-Asian cattle populations, with a focus on the understudied Thailand cattle, as a potential novel genetic resource for resilience and adaptability to climate change. A combination of medium and high-density Illumina Bovine SNP arrays was used, alongside a combination of genomic tools- the *PLINK* toolkit, STRUCTURE and TreeMix. These results revealed the population history and genetic structure of Asian breeds, validating previous research efforts which identify Thailand cattle as unique among other *Indicine* breeds, presenting an understudied genetic resource potential, requiring greater genetic management and further research.
We also investigated methods to increase efficiency of European dairy-cattle farms by reducing losses incurred by the prevalent parasite infection cryptosporidiosis. High-density Bovine SNP arrays was used for an association study with the aim of identifying selection signatures associated with *Cryptosporidium* infection. Using a combination of the *PLINK* toolkit, various R Packages and PANTHER Gene Ontology assessment, putative candidate genes associated with *Cryptosporidium* infection are discussed, and future research options are suggested. Putative genes include FMN2, TPM2 and TLN1 (novel), and CA9 and FGD4 (previously found to be directly/indirectly associated with *Cryptosporidium* infection).

Declaration:

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institutions of learning.

Acknowledgements:

I gratefully thank my supervisor, Dr Marta Farré-Belmonte, for accepting me on to this Msc-R project, and for her continued support throughout. She has provided me with technical aid both in the laboratory and with bioinformatics, and further to this, throughout the trials of research she has provided me with advice and oftentimes much-needed encouragement. Being a member of her research group has been a pleasure and has also taught me much about what it means to be a research scientist- to that end I would also like to thank all the members of the group- Dadu, Frances, Cristina, Sara, Carla and Dr Peter Ellis- and I would especially like to thank Corey for tirelessly answering my questions, and for his support.

I would also like to thank Dr Anastasios Tsaousis for the opportunity, and for sourcing and providing the samples used in this research, and thank you to Sumaiya, for providing the metadata.

Lastly, I would like to thank my friends for providing a pillar of love and laughter, a constant source of encouragement throughout my days.

Contents:

List of Figures

Figure 14. *Population graphs obtained with TreeMix. Holstein-Friesian of Bos taurus ancestry ('FRA') was denoted as the root (outgroup). For the 6 Asian subpopulations abbreviations see Table 1.* ***(A).*** *Shows the maximum likelihood tree of the 6 Asian subpopulations, rooted to the Hostein-Friesian outgroup.* ***(B).*** *The most likely number and structure of migration edges between the 6 Asian subpopulations and 1* Bos taurus *outgroup population.*

List of Tables.

Abstract

Livestock products have high densities of critical nutrients. As the world population increases, so too does the demand for animal products (such as milk), whilst climate change induced stressors are expected to intensify competition for resources, indicating that livestock systems must increase climatic resilience, alongside productivity and efficiency.

The genetic diversity of livestock breeds is shaped by evolutionary forces such as genetic drift, migration, selection and geographical separation. Modern livestock genetics have also been influenced by human-mediated selection. As a result, many highly specialized breeds are adapted to localised environments as well as having evolved to meet a variety of human needs. Both processes have left traces in the genome of domestic livestock species as genome-wide variants such as short-nucleotide polymorphisms, 'SNPs'. The growing availability of computationally efficient genomic tools means that selection signatures can be readily analysed to assess the genetic diversity and population structure in cattle breeds and among cattle populations.

This research utilised genetics in consideration of the growing need to safeguard livestock populations, by considering the need to increase efficiency as well as climate-change induced resilience, with a focus on cattle.

We investigated and analysed the population structure and diversity of South-Asian cattle populations, with a focus on the understudied Thailand cattle, as a potential novel genetic resource for resilience and adaptability to climate change. A combination of medium and high-density Illumina Bovine SNP arrays was used, alongside a combination of genomic tools- the *PLINK* toolkit, STRUCTURE and TreeMix. These results revealed the population history and genetic structure of Asian breeds, validating previous research efforts which identify Thailand cattle as unique among other *Indicine* breeds, presenting an understudied genetic resource potential, requiring greater genetic management and further research.
We also investigated methods to increase efficiency of European dairy-cattle farms by reducing losses incurred by the prevalent parasite infection cryptosporidiosis. High-density Bovine SNP arrays was used for an association study with the aim of identifying selection signatures associated with *Cryptosporidium* infection. Using a combination of the *PLINK* toolkit, various R Packages and PANTHER Gene Ontology assessment, putative candidate genes associated with *Cryptosporidium* infection are discussed, and future research options are suggested. Putative genes include FMN2, TPM2 and TLN1 (novel), and CA9 and FGD4 (previously found to be directly/indirectly associated with *Cryptosporidium* infection).

This research also provides suggestions for method improvement. For SNP Chip genotyping, nasal swabs were generally found to be preferable in terms of quality. The *Cryptosporidium* research was limited due to poor suitability of stool samples for SNP Chip genotyping as well as small scale data, which could be expanded upon for a full-scale GWA study in the future. Discussions on extending this research are also discussed, such as by investigating which *Cryptosporidium*-associated genes are up- or down- regulated, and whether associated with susceptibility or resistance. We also provide suggestions to improve Thailand cattle genetic analysis, such as developing programs for phenotypic performance recording, so genomic selection strategies can be applied in future research.

1.0 Introduction

The number of undernourished people is increasing: the prevalence of undernourishment increased from 8.4% in 2019 to 9.9% in 2020. This rise is prevalent in Africa, Asia, Latin America and the Caribbean[1]. Food insecurity presents a serious threat to public health, social sustainability, and political stability, a challenge further magnified as the world population is foreseen to increase, with current UN predictions projecting an increase to 9.7 billion people in 2050[2]. This pressure is accompanied by significant dietary structural change and demand; diets are shifting towards more high-caloric livestock products, including milk, and away from staples such as roots and tubers, a shift predominantly reflecting economic gains of developing countries[3].

At the same time, climate change induced stressors are expected to intensify competition for resources. This presents a dual problem for livestock systems, which must increase climate-change related resilience, alongside productivity and efficiency. Advances in genomic tools means that selection at the molecular level is becoming more accessible to farmers and now has the potential to expedite both pure- and crossbreeding programmes for breed improvements. Considering that cattle are the most common large livestock species in the world- the global population size of which is approximately 1400 million animals[4]- there is need to safeguard this livestock species against the effects of climate change, as well as increase in efficiency to feed a growing population.

1.1 Introduction to Modern Cattle.

To fully appreciate the genetics of modern-day domesticated cattle, a review of their historical evolution is justified.

Modern domesticated cattle comprise two extant species, *Bos taurus* (taurine) and *Bos indicus* (indicine, otherwise known as 'zebu'). Both species originated from the extinct *Bos primigenus* (auroch)[5]. Taurine cattle are thought to originate from the Fertile Crescent, and Indicine cattle were domesticated from the Indus valley, likely from the Indian auroch subspecies *B. primigenius namadicus*[6], with both domestication events occurring between ~8,000 and ~10,000 years ago[4]. The divergence between the taurine and indicine lineage has been dated ~250,000 years ago according to mitochondrial DNA haplotype analysis[6] and furthers the argument for cattle having two independent domestication sites (although there may be a third[8]; see *Figure 1*).

The domestication of cattle aided agricultural development, transforming migratory human hunter-gatherer populations into settling farming societies[7]. Since then, as humans have colonised the Earth, cattle have also spread across the globe (*Figure 1*), with farmers selecting animals for desirable characteristics and establishing breeds. Until recently, this process was slow and enabled animals to adapt to local environmental selective pressures. The domestication and artificial selection led to the development of genetically divergent cattle breeds (or hybrids) that exhibit specific genetic diversity patterns and population structure.

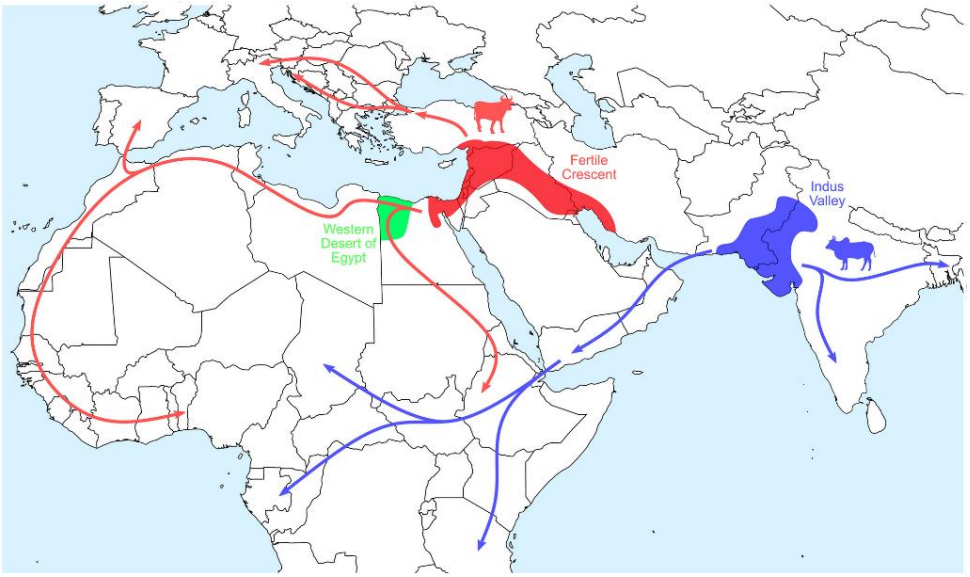


Figure 1. *Hypothesized main domestication sites and migration routes of taurine and indicine cattle, including a postulated third domestication site in Egypt. Source: Pitt, D. et al. (2019)[8].*

On both local- and industrial- scales, cattle are a critically important daily source of food and nutrition. Due to advancements in hereditary understanding, breeding methods have developed rapidly. The acceleration in artificial selection has driven the expansion and value of agricultural sector[5]. Thus, to increase or maintain breed purity, the gene-flow between genetically varied cattle has been reduced. This has resulted in high-yield and high-quality commercialised breeds, such as the Holstein-Friesian breed (specialised dairy cattle). To highlight the intense improvement of the breed, despite the UK dairy cow population having reduced by 28% from 1996 to 2020, the UK also produced 15.3 billion litres of milk in 2020, the highest annual figure since 1990[9].
However, because of breed improvements on this scale, recent breed improvement strategies are generally counter to genetic variability[10] and as will be discussed, threaten loss of valuable genetic resources.

Considering this diversity, this research will investigate both the importance of cattle to global food security and methods to continue efficiency improvement (1.3), as well as the importance of localised breeds to maintaining genetic resources (1.2).

1.2 Maintaining genetic resources and improving genetic health of local breeds.

The development of the breed concept, whilst improving global food security as discussed in 1.1, also presents a threat to available genetic resources which may be otherwise useful to ensure adaptation potential for livestock breeds under warming conditions- this is especially true for breeds found in warmer climates. However, past studies have mostly addressed breeds from developed countries, where climate-control is widely practiced, highlighting the need to genetically characterise locally adapted breeds, and relatively quickly- to meet the demand of growing food insecurity, local farmers are replacing, or crossbreeding, native breeds with commercially available breeds.

A total of 990 cattle breeds are known throughout the world. Only 93 are considered transboundary breeds (for example, the Holstein-Friesian) breed, whilst the remaining 897 are classified as local or indigenous breeds, highlighting a vast number of potential genetic resources. Among these, 258 breeds are reported to be present throughout Asia. An FAO watchlist found that 11% of these Asian breeds are at currently classified as *at risk*, and 38% of breeds are classified as unknown[11]. Hence, Asian breeds- especially those not yet genetically characterised, such as Thailand cattle- present a potential novel source of genetic resource for adaptability to warming conditions under climate change.

The first step in this process requires additional information relating to the evolutionary history and diversity of breeds in Southeast Asia[12,13]. A recent worldwide study of bovines, including Asian indicine breeds, demonstrated a complex history for Asian diversity; there are three possible ancestries for the cattle in the area, taurine, indicine and the locally domesticated *Bos javanicus* ('Bali')[14].

Historically, cattle in Thailand were used primarily for draft and meat production, but from the 1950s, cooperation between the Thailand government and the UN led to the initiation of a school milk program to improve the nutritional status of local communities[15]. Cross-breeding occurred between imported purebred Holstein cattle (*B.taurus*) to improve milk quality and yield in local livestock[16,17]; currently, 95% pure Holstein trait is retained in less than 75% of the Thailand cattle population[18,19]. This indicates that Thailand cattle have undergone (re)selection for the local environment over the course of 60 years, developing resistance to tropical diseases and external parasites, and able to be sustained on low-quality roughage and grasses following traditional and localised methods of animal husbandry, representing useful traits for warming conditions under climate change.

Despite the importance of Thailand cattle, genetic improvement and management of this population is limited, since farmers lack records concerning cattle pedigree and rely on personal experience to make breeding decisions[20]. Knowledge of the genetic diversity is required to facilitate effective management, and phenotype/performance scoring would aid genomic selection in the future.

Improving knowledge of the genetic diversity and population structure of Thailand cattle, in their relation to other Asia breeds (*Figure 2* and *Table 1*), will therefore provide a rational basis for the populations overall genetic health and inform breeding strategies. This hence justifies investigating the Thailand cattle, with the eventual aim of conserving the genetic integrity of this locally adapted

breed, as both a source of nutritional and high-caloric, sustainable and local food production and as a genetic source of climate change resistance.

Figure 2: *Geographic location of the Asian cattle breeds/populations used in this dataset. Thailand cattle represent an understudied population of cattle with previous research on Asian cattle seldom including Thailand analysis.*
Map created with online tool 'mapchart.net'

Table 1*: The nomenclature (including abbreviations) of the Asian Cattle Breeds and populations used in this research, excluding the Thailand cattle.*

Lab ID.	Abbreviation.	Full breed name.	Region of origin.
105	WAG	Wagyu	Japan.
106	HAN	Hanwoo	Korea.
227	MON	Mongolian	Mongolia.
618	BRE	Brebes	Java (Indonesia).
619	MAD	Madura	Madura Island (Indonesia).

1.3 Improving the efficiency of industrialised dairy cattle by reducing parasite-incurred losses.

The value of dairy cattle to this growing agricultural industry is already at an all-time high. Total world milk production comprises 81% cow milk and is currently valued at £507.16 billion, and it is expected to keep growing by an increase of 1.7% p.a over the next decade, faster than most other main agricultural commodities[21]. In terms of food security, milk is the third biggest supplier of protein and the fifth largest provider of calories, presenting an important source of high nutrition in the fight against undernourishment[22].

If the demand for nourishment continues as projected, by 2050 there would be a requirement of 120% more water, 42% more cropland, and a loss of 14% more forest to supply sufficient agricultural land[23,24]. With this scenario, there would be ~70% increase in global greenhouse gas emissions. Hence, to meet this challenge, there is a need for sustainable intensification of agriculture on land that is already available ('sustainable intensification')[25], justifying the need to investigate methods to improve efficiency and production yield. The United States Department of Agriculture reported that, during 2017, ~1.9 million calves were lost to nonpredator causes, with respiratory problems accounting for the highest percentage of losses (26.9%), followed by digestive problems (15.4%)[26]. Total combined cattle and calf death losses were valued at $3.87 billion in 2017. This forecast justifies an in-depth investigation into prevalent cattle diseases, which contribute to losses via increased mortality, reduced productivity, control costs, loss in trade, decreased market value, and thus greater food insecurity.

A major disease of European dairy cattle is cryptosporidiosis, which causes a decrease in their productivity. As part of tackling this problem, this research investigated dairy cattle infected with the parasite *Cryptosporidium* in Europe, performing an association analysis comparing the genetics of *Cryptosporidium* infected animals with the genetics of non-infected animals. In the long term, selective breeding for parasite resistance in combination with other integrated control methods is considered an alternative means of parasite control; resistance to infection by various endo-parasites (notably not *Cryptosporidium*) in beef and dairy cattle was found to be associated with various QTL's[27], and another study in sheep found significant genetic variation between resistant- and non-resistant- animals[28].

1.3a Review of *Cryptosporidium* parasite life cycle and the effects of cryptosporidiosis diseases.

Cryptosporidium is a globally distributed protozoan parasite which has been found in many vertebrates, including cows and humans. *Cryptosporidium* causes cryptosporidiosis, an enteric infection leading to gastrointestinal (GI) illness such as severe diarrhoea. It can be fatal to groups which are especially vulnerable to the disease, including pre-weaned animals.

Cryptosporidium infections occur after ingestion of oocysts through the faecal–oral route. Understanding the life cycle of *Cryptosporidium* in the bovine host is integral to understanding its disease burden[29]. The *Cryptosporidium* zygote will produce either thin-walled oocysts, which re-infect the host, or thick-walled oocysts, which are shed in stool and are immediately highly infective. This is further amplified considering that infected cattle individuals can shed up to ~100 thick-walled oocysts per passage, which remain viable in the environment for up to 2 months after cessation from the host[29, 30.] The thick-walled oocysts are highly resistant to chemical treatment, including

water chlorination[31]. Hence, management of *Cryptosporidium* is challenging; prevalence of cryptosporidium in stool samples of European cattle herds were reported to range from 13% to 100%[32].

Hence for cattle, and especially for intensive cattle farming systems (such as industrialised dairy cattle farms), cryptosporidiosis is considered globally endemic[30]. The disease has significant economic burden, with costs associated with veterinary diagnosis and medication, animal rearing and supplemental nutrition nearing 100-200 GBP per *Cryptosporidium*-infected calf[33]. Further, the severe Cryptosporidiosis is associated with reduced long-term growth rate in calves, causing economic losses (see *Figure 3*).

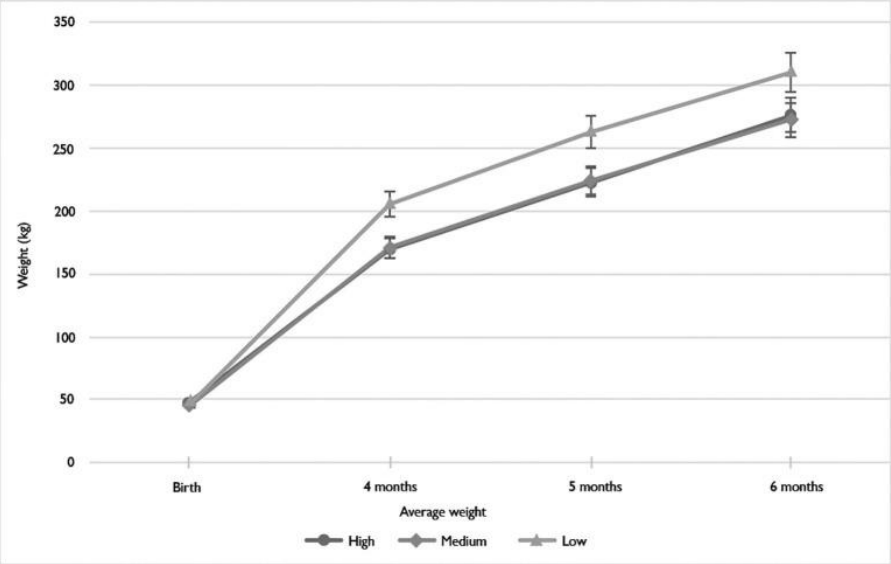


Figure 3. *Average weights of calves over a 6-month period based on their cryptosporidiosis severity level (high, medium and low). Error bars represent 95% confidence interval of the mean. For further information regarding this graph, the source: Shaw et al (2020).*

Few tools are available to combat cryptosporidiosis. Aside from the general diarrheal-aid treatments, only one drug is considered effective and only suitable for human immunocompetent hosts (nitazoxanide)[34]. Hence, for improving the cattle industry, genetic improvement presents a feasible strategy in combatting cryptosporidiosis. A quick and economically viable method of identifying underlying genetic factors associated with *Cryptosporidium* infection is to use an association study to identify *Cryptosporidium*-infection associated short nucleotide polymorphisms ('SNPs), which may be present in, or nearby genes putatively associated with *Cryptosporidium* infection.

1.4 Detecting genomic differences.

When a favourable mutation occurs within a population under directional selection, the frequency of the favourable allele is likely to increase over time. Nucleotides adjacent to the favourable mutation also tend to increase in frequency, a process known as “hitch-hiking”; that is, when one gene is undergoing “selective sweep” within a population, nearby polymorphisms that are in linkage disequilibrium also tend to change their allele frequencies[35]. This process leads to identifiable “selection signatures”, as characterised by nucleotide distributions about the favourable mutation that differ statistically from that expected by Hardy-Weinberg Equilibrium[36].

Detection of selection signatures increase our understanding of the evolution and biology underlying a given phenotype. Recent developments have provided tools which further the detection of these selection signatures, such as genotyping. Genotyping is the process of determining differences in the genotype of an individual, firstly by examining the individual's DNA sequence using biological assays and comparing it to a priorly sequenced reference genotype.

SNP genotyping is a type of genotyping which measures and identifies *single-nucleotide polymorphisms* between members of a species. SNPs are a single base pair mutation at a specific locus that is conserved during evolution and are useful markers for association studies.

Currently, two main approaches can be used to detect SNPs at large scale: SNP Chips or re-sequencing.

SNP Chips are DNA microarrays that identify SNPs at pre-specified locations across the genome. SNP Chips contain a set number of immobilized allele-specific oligonucleotide probes containing the pre-specified SNP library. The target DNA fragmented, the sequences amplified via PCR and then labelled with fluorescent dyes and hybridised with the SNP Chip. After being washed the fluorescence is scanned and processed via a detection system that records and interprets the hybridization signal. SNP Chips have proven useful for assessing common genetic variation within a population (useful for inter-population studies of the same species), as well as predisposition to complex multifactorial diseases (for example, susceptibility to parasitic infection). However, since SNP assays contain pre-specified SNPs, there may be ascertainment bias if the SNP array or DNA sample is not representative.

Re-sequencing uses restriction enzymes to digest genomic DNA which is then ligated to a barcode adapter and amplified via PCR. Libraries are then sequenced on a single lane of flow cells, using Next Generation Sequencing methods. This approach can process multiple DNA samples at once, and processes complex genotypes. It is especially efficient for high-throughput of high-diversity, large genome species. This approach also allows for the scoring of rare SNPs (for example, that may not be included in a SNP Chip) because of low polymorphism information content, which may be important contributors to adaptation.

In general, SNP Chips are a reliable and robust platform for use in breeding programmes and diversity research, with many cattle SNP Chips readily available. Re-sequencing -scored SNPs have the capacity to deliver many markers and are especially useful for organisms without a reference genome but may also return considerable amounts of missing values and unknown marker positions. With these considerations, SNP Chips are considered more reliable for use in cattle genomics, especially for *Bos taurus* breeds.

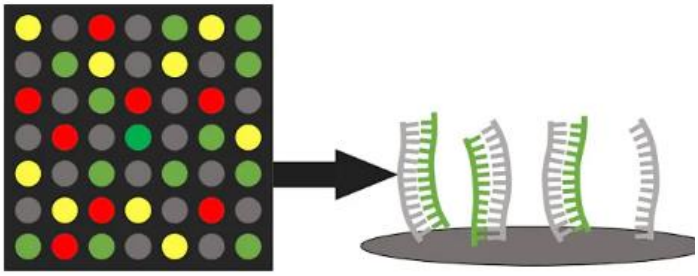


Figure 4: *Example of SNP Chip (DNA microarray) method. The hybridisation signal is interpreted via fluorescence, for example, whether the individual genotyped is homozygous, heterozygous, or contains the minor allele.*

For association-based studies, the *'indirect'* approach can be used to identify *regions* under selection, by comparing the distribution of allelic frequencies between a group with the phenotypic trait of interest, with a control group without the phenotypic trait of interest. Therefore, this approach seeks to identify association between a particular genomic region and the phenotype, whether the variants themselves have a direct functional/known effect. The search for causative variants is then constrained to associated genomic regions.

On the other hand, for population studies, allelic frequency distribution across populations is useful to characterise the genetics, and therefore, identify populations and understand gene flow between them, the genetic health of the population (informed via inbreeding and heterozygosity statistics), and further to this, understanding the regions under selection associated with local adaptive traits.

Post-association analyses such as gene ontology (GO) term analyses enables a better understanding of the molecular mechanisms underlying the trait-associated-gene.

1.5 Hypothesis.

We hypothesise that there will be distinct population genetic differences (high variability) between the North vs the South Asian cattle breeds, with southernmost (Indonesian) breeds having low intra-breed variability but high inter-breed variability due to relative geographical isolation. Secondly, we hypothesise that by comparing infected vs non-infected cattle, we will identify genomic regions associated with *Cryptosporidium* infection, considering the results of prior research finding other endo-parasitic infection resistance to have a genetic basis.

1.6 Overview of research aims.

To test our hypothesis, we divided our research into three main aims:

1. To optimise DNA extraction protocols and determine the most suitable sample type for use in SNP Chip Genotyping.
2. To investigate and discuss the genetic diversity of Asian Cattle subpopulations, with a focus on the understudied Thailand population, and inform areas for improvement in local cattle farming practices, with aim of sustaining genetic resources for future use.
3. To reveal putative candidate genes, and/or putative genomic regions, associated with Cryptosporidium infection and lay the foundations for further research in this area.

2.0 Materials and Methods.

2.1 Sampling and DNA Extraction.

DNA was extracted from 30 cows and calves of varying sample types, summarised in *Table 2*. Different methods were used for the extraction, depending on sample type; phenol-chloroform extraction for nasal swabs, QIAamp Fast Stool Mini Kit for stool samples and QIAamp DNA Blood Mini Kit for blood samples. The quantification, yield and purity of the DNA was measured on a Nanodrop.

Additional details of the European samples (Netherlands, Cyprus and France) are available in *Appendix table 1.*

Table 2. *Summary of the sample types and DNA extraction methods used.*

Country Origin	Breed	No. Individuals	Sample Type	DNA Extraction method	*Cryptosporidium* status
Thailand	Zebu	5	Nasal swab.	Phenol-chloroform extraction.	N/A
Netherlands	Holstein-Friesian	10	Stool.	QIAamp Fast DNA Stool Mini Kit.	5 infected, 5 non-infected.
Cyprus	Holstein-Friesian	10	Stool.	QIAamp Fast DNA Stool Mini Kit.	5 infected, 5 non-infected.
France	Holstein-Friesian	5	Blood.	QIAamp DNA Blood Mini Kit.	Non-infected.

For the European *Cryptosporidium* analysis, stool samples were collected from 20 Holstein-Friesian calves from the Netherlands and Cyprus, comprising 10 *Cryptosporidium* positive calf individuals and 10 *Cryptosporidium* negative individuals.

Parasite infection was determined via identification of oocysts in stool under bright-field microscopy. Stool sample type and health was also classified according to the Bristol Stool Score *(Appendix Table 1*). Stool samples were stored in -20°C. DNA. DNA was extracted and purified according to the QIAamp Fast DNA Stool Mini kit manufacture protocols and eluted in 200µl. The extracted DNA was precipitated using 3M Na-Acetate (pH 5.1) and 100% ethanol and incubated for ~24hrs at -20°C, followed by 2 washes with 70% ethanol. DNA was resuspended in 50µl ddH_2O.

To ensure the DNA extracted contained cattle DNA (as opposed to, for example, microbial), a PCR was performed. Primers were designed to amplify a region of 261bp length of the bovine transferrin receptor 2 (TFR2) gene.
The PCR reaction mix contained 1x PCR buffer, 200 µM of each dNTP, 0.5 µM forward primer, 0.5 µM reverse primer, and 2.5 units/reaction 'Qiagen HotStarTaq DNA Polymerase'. A 15-minute denaturation stage was performed at 95°C, followed by 35 cycles of annealing and elongation (94°C for 1 min, 50°C for 1 min, and 72°C for 1 min), ending with 10 mins at 72°C.
Amplification results were visualised with 1% agarose gel electrophoresis under a Syngene Gel Doc.

Blood samples were collected from 5 Holstein-Friesian individuals from France (ages of 2 calves, 1 cow, 3 unknown). The blood collected via venepuncture and stored in Vacutainer EDTA 3ml tubes. The samples were stored in -80°C and thawed to room temperature for DNA extraction and

purification according to the QIAamp DNA Blood Mini Kit manufacture protocols. DNA was eluted in 50µl ddH_2O.

For the Asian *indicine* ('Zebu') diversity analysis, nasal swab samples were collected from 5 *Bos indicus* ('Zebu')-type cattle individuals from Thailand. Samples were stored in -20°C and DNA was extracted and purified using Phenol-Chloroform extraction methods, as adapted and optimised from previous protocols[37].

An equal volume of phenol:chloroform:isomyl-alcohol was added to the nasal swab and placed on a rocking platform for 30 mins. Samples were then centrifuged at 15,000x g for 5 minutes at room temperature and the upper aqueous phase was transferred to a clean Eppendorf tube. DNA was precipitated by adding an equal volume of isopropanol and centrifuged for 15 mins at 13,000x g at 4°C. Isopropanol was removed, the pellet was rinsed with 70% ethanol, and dried at room temperature. The pellet was dissolved in 50µl ddH_2O.

All extracted and purified DNA samples were quantified for yield and quality on a Nanodrop. Samples were made to 35µl of 20ng/µl (yield of 700ng) for genotyping. DNA samples were stored in Eppendorf LoBind Tubes, sealed with parafilm, and sent to NeoGene in wet ice.

Samples were genotyped with GeneSeek® Genomic Profiler™ Bovine 100K (Illumina Inc., San Diego, CA, USA), mapped to the ARS-UCD1.2 assembly.

2.2 SNP Data Analysis.

Quality of the raw SNP data was examined using PLINK[38] (Whole genome association analysis toolset) via RStudio.

Analysis of the raw SNP data quality was performed using summary statistics available in PLINK: --freq modifier (Minor Allele Frequency report), the --missing modifier (sample-based and variant-based missing data reports), and –het modifier (Inbreeding via observed and expected autosomal homozygous genotype counts). Due to lack of data relating to *Cryptosporidium* infection status, the French blood samples were excluded from further analysis.

To filter the raw SNP data, the following PLINK modifiers were used: --geno, --mind, --maf, --hwe. For each modifier, the threshold was altered according to raw sample quality results.
Two quality-control scenarios were used accordingly, given in *Table 3*. The external dataset was filtered according to optimal scenario QC1, and the SNP data derived from stool and nasal swabs (*Cryptosporidium*- and Thailand- samples) according to less-than-optimal (less conservative) scenario QC2.

***Table 3**. Overview of modifier effect thresholds used in each quality control scenario, as modified according to the raw sample quality outcomes.*

		Scenario	
		QC1 (Optimal)	**QC2**
Modifier in effect	**-- mind**	0.1	0.4
	-- geno	0.1	0.2
	--maf	0.00005	0.05
	-- hwe	0.000001	0.000001

The script used for subsequent PLINK/R analysis are given in *Appendix 2.1* (. The full methods and scripts for implementation of PLINK files into the TreeMix software is given in *Appendix 2.2.)*

With the data filtered, further downstream analysis could be performed. The methods now split according to European *Cryptosporidium* analysis (2.3) and Asian Cattle Diversity analysis (2.4).

2.3 European *Cryptosporidium* Analysis (*Bos taurus*).

The European *Cryptosporidium* dataset remaining after quality control (scenario QC2) was used for further downstream analysis, clustered according to *Cryptosporidium* Positive individuals ('POS') and *Cryptosporidium* Negative individuals ('NEG').

The POS and NEG PLINK files were merged using the --merge PLINK function. Resulting duplicate SNPs were removed by generating a .dupvar file (duplicate-position-and-alleles variant report) using the PLINK function --list-duplicate-vars, to then be excluded using the --exclude function. The PLINK function --fst was used to generate a fixation index report (.FST file) for visualisation into a pairwise FST Manhattan plot per SNP position per chromosome, using ggplot package.

Regions of Interest (RoI) were denoted as having > 0.45 FST. Of these, genes of Interest (GoI) were extracted with the biomaRt package within windows of +/- 50,000bp of the RoI, with the ENSEMBL dataset given as 'btaurus'. Genes were returned and listed according to "ensembl_gene_id", "external_gene_name", and "gene_biotype".

The list of ENSEMBL gene ID's was classified according to the Panther Gene Ontology[39] classification system. These genes were analysed according to 'Functional classification viewed in gene list' and 'Functional classification viewed in graphic charts' options.

2.4 Asian Cattle Genetic Diversity (*Bos indicus*).

To identify the genetic origin of Thailand samples, we used an additional dataset including 5 additional breeds (41 individuals). This allowed the addition of 9 individuals of the Brebes breed and 7 individuals of the Madura breed (Indonesian origin) which, in addition to our 5 Thailand cattle, comprised the South Asian Cattle breeds. Also included from this additional dataset was 12 individuals of the Wagyu breed (Japanese origin), 5 individuals of the Mongolian breed (Mongolian origin) and 8 individuals of the Hanwoo breed (Korean origin), which comprised the Northern Asian Cattle breeds[40].

The additional dataset, mapped to Illumina BovineSNP50 BeadChip, was updated from the UMD3.1 Assembly to the ARS-UCD1.2 assembly in its positions and SNP names, using the NCBI Genome Remapping Service[41]. For filtering, a subset was created for the additional dataset breeds (BRE, MAD, HAN, WAG, MON) and a separate subset for the Thailand cattle, based on data quality assessment. Hence, the external dataset was filtered according to QC1, and the THA cattle according to QC2.

After filtering, the additional dataset and the Thailand dataset were merged using the --merge function in PLINK, giving 44 individuals comprising 6 Asian breeds (for merging datasets based on common SNPs, see script in *Appendix 2.1*). Any resulting duplicate SNPs were removed by generating a .dupvar file (duplicate-position-and-alleles variant report) using the PLINK function --list-duplicate-

vars, to then be excluded using the --exclude function. The resulting merged and filtered dataset was used in downstream analysis.

2.4 (a) Population Structure Analysis.

To study the population structure of the Asian cattle breeds we used four methods: FST, principal component analysis (PCA), STRUCTURE and TreeMix.

An .fst file was generated using the PLINK --fst function with pairwise FST results given for Thailand Cattle vs the DRYAD Asian cattle breeds.

A PCA was performed based on the 6 breeds, with the genetic distance between the 44 individuals calculated using the PLINK function --distance-matrix and the PCA plot generated using the cmdscale function, according to eigenvalues and eigenvector properties. Principle components 1 and 2 were visualised using the ggplot package.

PLINK files were converted into files suitable input into STRUCTURE software using the PLINK --recode function (flag 'structure').

STRUCTURE[42] analyses differences in the distribution of genetic variants among populations with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation. Runs of STRUCTURE were carried out assuming between 1 and 8 groups (K), with a burnin period of 500 cycles and 1,000 data collection Markov chain Monte Carlo (MCMC) cycles. Three iterations were performed for each value of K. STRUCTURE Harvester[43] was used to determine the best K value. STRUCTURE Harvester utilises the "Evanno" method[44], i.e it calculates an *ad hoc* statistic based on second-order rate of change in likelihood weighted by the standard deviation. The Delta (K) captures the genetic divergence between significantly distinct populations, and the L(K) value predicts the K value with the highest probability.

Finally, to detect the historical relationships among populations we used TreeMix[45]. This is a method for inferring the patterns of population splits and mixtures in the history of a set of populations, by generating a graphical representation of both population splits and migration events.

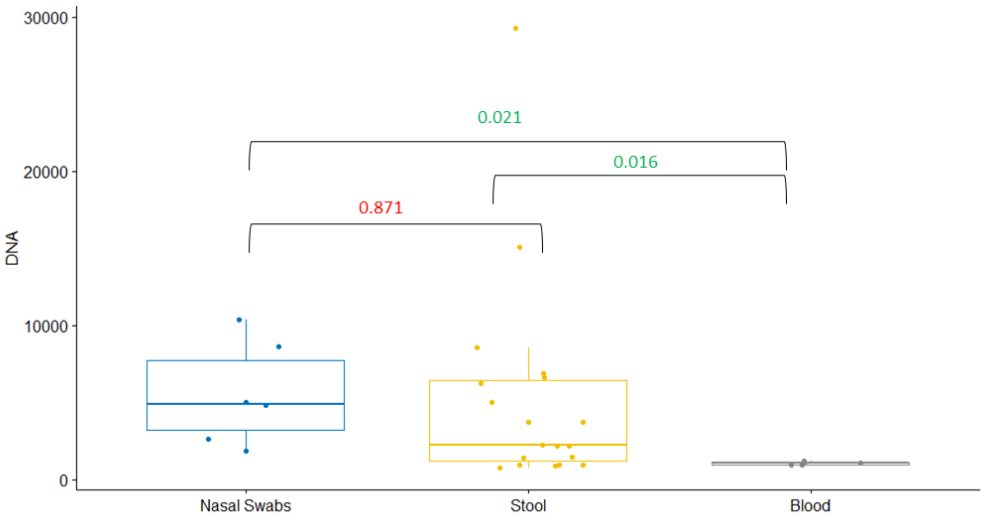
The script used for generating TreeMix results is given in *Appendix 2.2*. Migration edges of 0-5 were ran, using the Holstein-Friesian (*Bos taurus*) breed as the outgroup. A bootstrap replicate was applied by resampling blocks of 500 SNPs. The OptM[46] programme was used to estimate the optimal number of migration edges on population trees. OptM utilises the "Evanno" method.

3.0 Results.

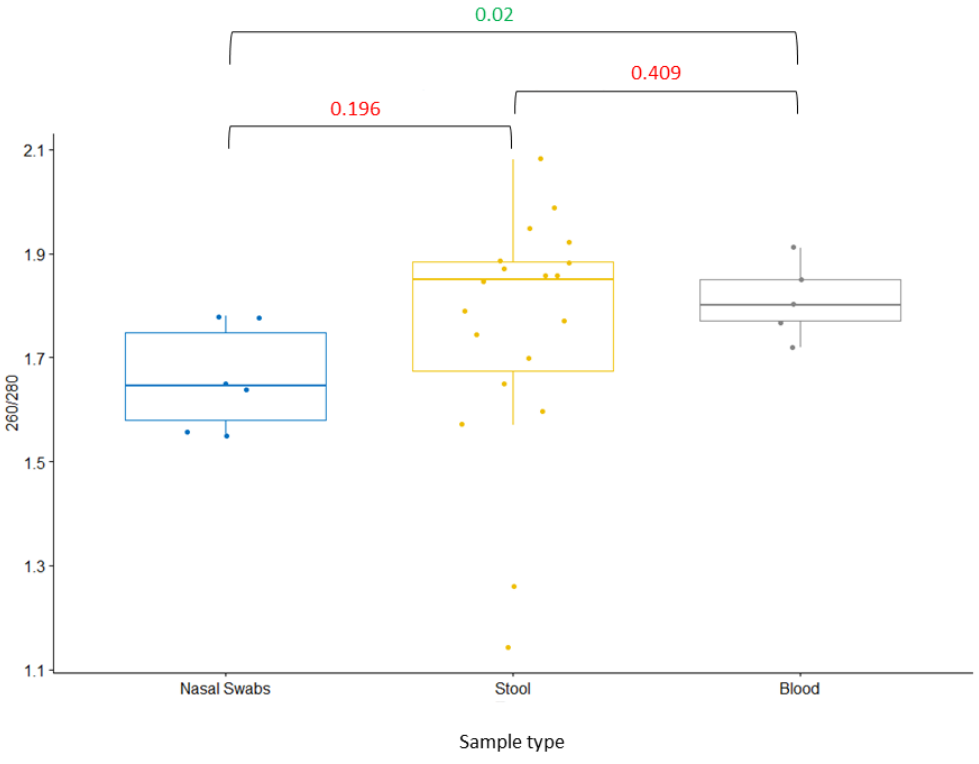
This research aimed to investigate the use of genetics in preserving cattle agriculture on both industrial- and localised- level farming, by considering methods to improve efficiency as well as climate-change induced resilience. To address the effectiveness of applied method, DNA yield, quality and %Missing call-rate was compared between sample types (nasal swabs, blood and stool) using a two-tailed t-test. To improve the efficiency of industrial dairy cattle farms, we investigated the genetic regions associated with *Cryptosporidium* infection by conducting an association study, revealing genomic regions and nearby genes for functional-classification analyses. To consider the potential of Thailand cattle as a genetic resource for climate-change resilience, we investigated the relation of Thailand cattle to other Asian breeds in terms of population difference, and then the evolutionary relationships between them.

3.1 DNA Quality according to sample type.

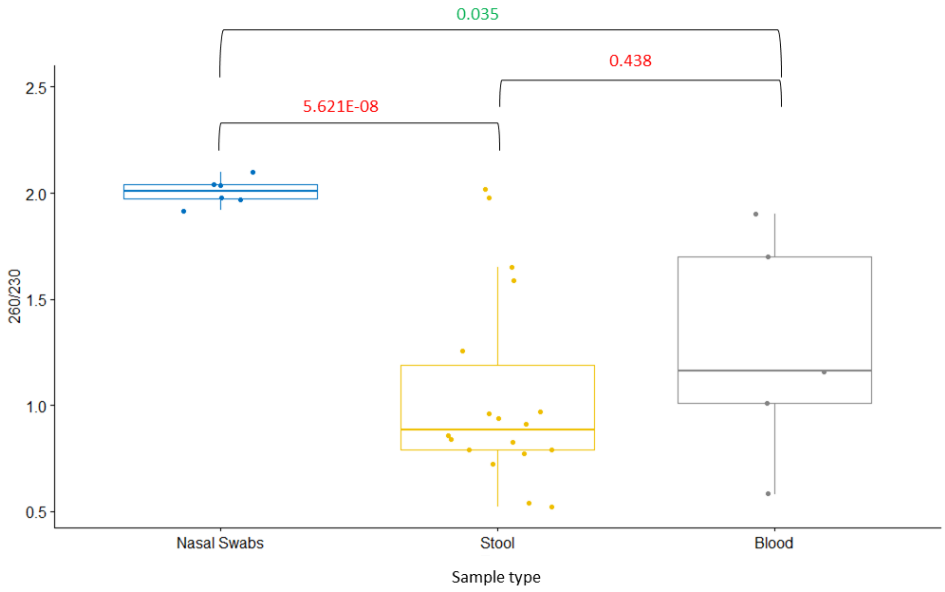
Three sample types (stool, nasal swabs and blood) and the subsequent three methods of DNA extraction were used to assess DNA quality and suitability for SNPChip genotyping. DNA quantity and quality measurements, as determined on a nanodrop, are shown in *Figure 5* with pairwise two-tailed t-tests yield and quality.



(A) *DNA Yield (ng).*



(B) *260/280 measurement.*



***(C)** 260/230 measurement.*

***Figure 5.** Box plots of Nanodrop measurements per sample type, including the P-value calculated via a two-tailed t-test. **(A)** DNA yield (ng), **(B)** absorbance ratios of 260:280 and **(C)** absorbance ratios of 260:230. Green values are considered statistically significant (P-value:<0.05). Note: stool sample Np25 was excluded as an outlier for DNA yield (having 162,750ng).*

Stool samples ranged in DNA yield from 745ng to 162,750ng, with an average of 13,096.5ng; contrariwise, DNA yield obtained from blood samples ranged from 935ng to 1,220ng, with an average of 1059ng, indicating a greater consistency in DNA extraction yield obtained from blood samples. Nasal swabs ranged in DNA yield from 1,845ng to 10,405ng, with an average of 5,561.66ng.

The DNA quality varied between sample types. DNA purity for SNPChip genotyping, measured as absorbance at 260/280 or 260/230, is optimally ~1.8 or ~2.0, respectively. Nasal swab samples extracted via Phenol-chloroform extraction produced the better-quality DNA (average 1.66 for 260/280 and 2 for 260/230) and stool samples via QIAamp Fast DNA Stool Mini kit produced the lowest quality DNA, with average quality reading 1.68 for 260/280 and 0.98 for 260/230. Blood samples also had low 260/230 readings, with an average of 1.27.

Yield (ng) and absorbance ratios of 260/280 and 260/230 were significantly different between the nasal swabs and blood samples. The 260/230 ratio was significantly higher between nasal swabs and the stool samples, however the 260/280 ratio was not significant between nasal swabs and stool samples. The blood samples had significantly lower DNA yield when compared to stool samples. This indicates that out of the samples investigated, nasal swab samples produced the better-quality DNA, though not necessarily the highest in yield.

Considering the large range and average yield of DNA for the stool samples, a PCR was performed. This specifically detected and confirmed the presence of cattle DNA by amplifying a region of 261bp of the bovine transferrin receptor 2 (TFR2) gene (*Figure 6*).

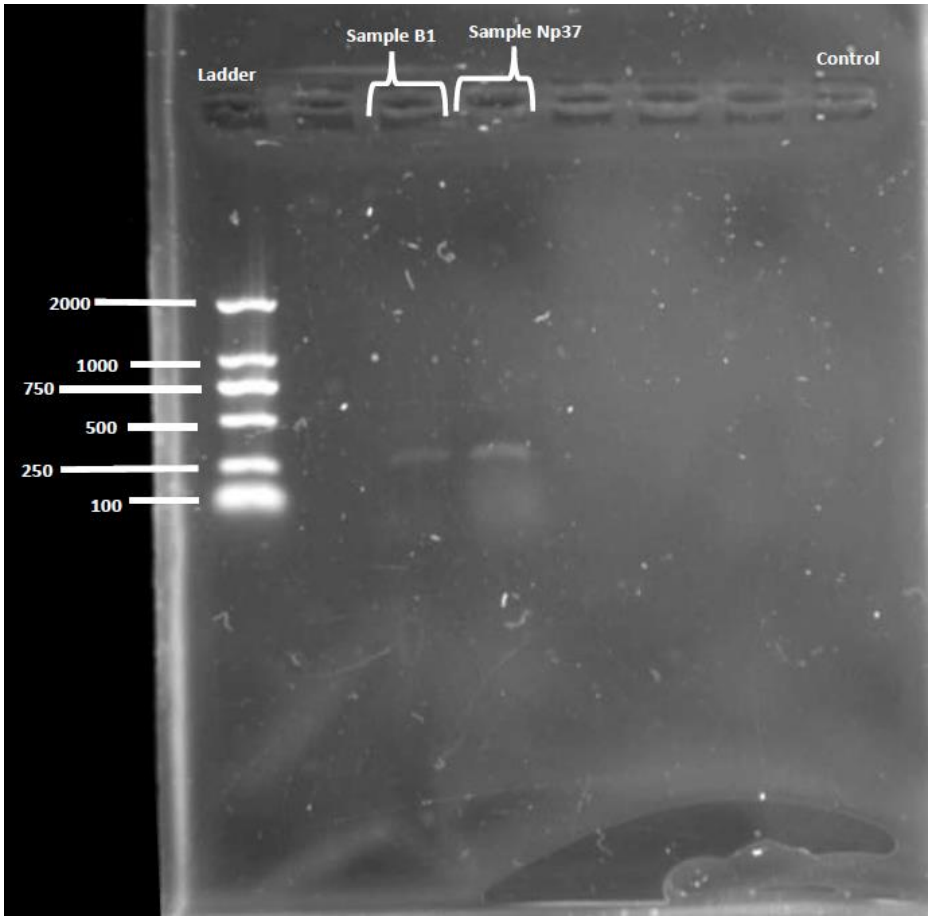


***Figure 6**. PCR performed on random cattle stool sample (Sample B1) and stool sample Netherlands Cryptosporidium Positive (Np37). The primers were designed to amplify a region of 261bp bovine transferrin receptor 2 (TFR2) gene, as visualised.*

3.2 Suitability of sample type for SNP Chip sequencing.

A total of 30 individuals were sent for genotyping in the GeneSeek® Genomic Profiler™ Bovine 100K, as mapped to the ARS-UCD1.2 assembly. Raw SNPs were obtained, and we then proceeded to quality check the dataset. First, SNPs were filtered according to QC2 *(Table 3*). Even with the QC2 less-than-optimal filters, 60.82% of SNPs were lost due to poor quality, including 54,675 SNPs removed due to missing genotype data and 275 SNPs removed due to minor allele frequencies.

***Table 4**. QC2 filtering of extracted DNA samples.*

Type	Before filtering	After filtering	Samples
Individuals	30	22	(5 CYN, 2 CYP, 4 FRA, 3 NEN, 4 NEP, 4 THA)
SNPs	90,349	35,399	

After filtering, blood samples had the lowest Missing Call Rate (Figure 7), with an average of 0.000784% missingness. Nasal swab samples had the highest average missingness of 0.0821%, a significant difference compared to blood samples, but not to stool samples, which had an average of 0.0627%.

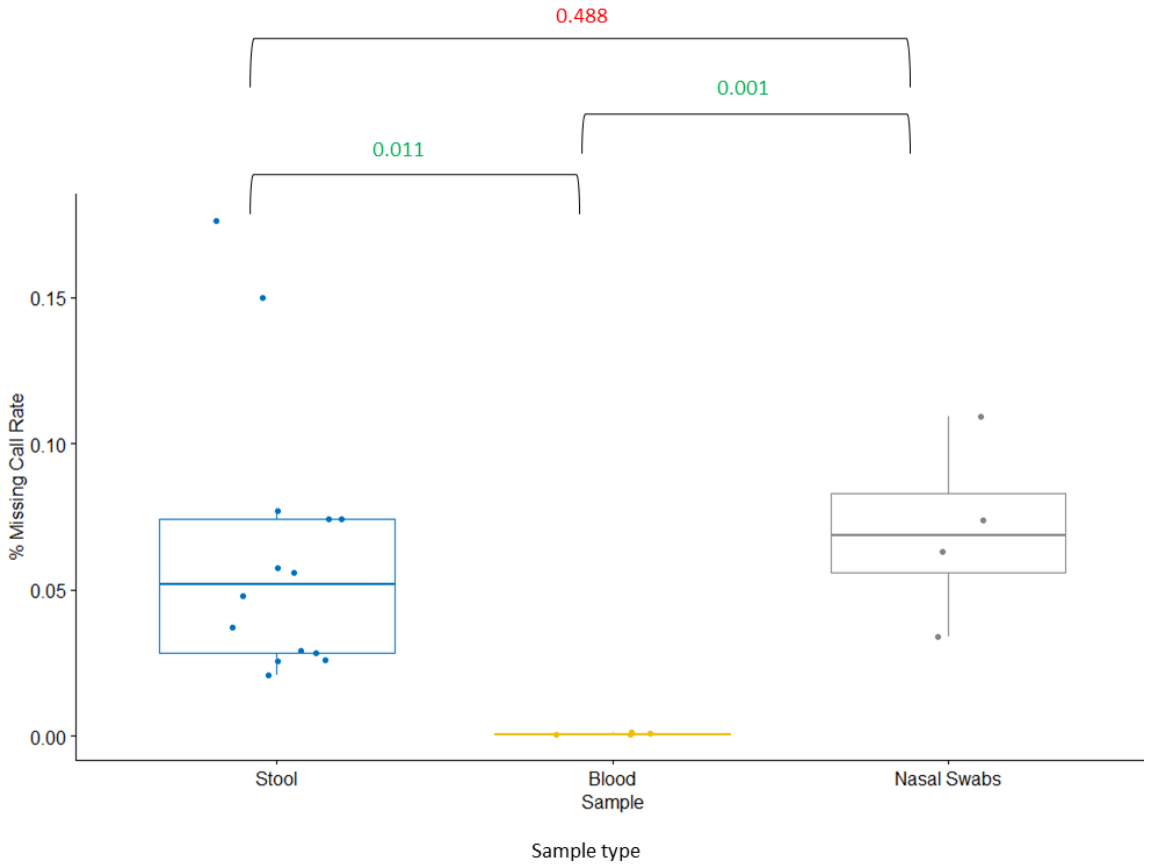


***Figure 7.** Box plots of %Missing Call Rate of SNPs per sample type, with P-values derived from a two-tailed t-test. Green values are considered statistically significant (P-value:<0.05).*

3.3 European *Cryptosporidium* Analysis.

To identify genomic regions associated to *Cryptosporidium* resistance in dairy cattle, we first grouped individuals by their *Cryptosporidium* status: *Cryptosporidium* Positive individuals ('POS') and *Cryptosporidium* Negative individuals ('NEG'), regardless of country-of-origin.

This dataset, initially comprising 10 *Cryptosporidium* positive individuals and 10 *Cryptosporidium* negative individuals with 90,349 SNPs were filtered according to QC2, with the resulting merged dataset (minus duplicate positions) comprising 67,970 SNPs and 17 individuals (10 NEG and 7 POS) passing the quality control for use in downstream analysis.

As expected, because all individuals are from the same breed, the PCA results revealed a unique population. There were four outliers from the main cluster which are likely due to varying sample DNA quality.

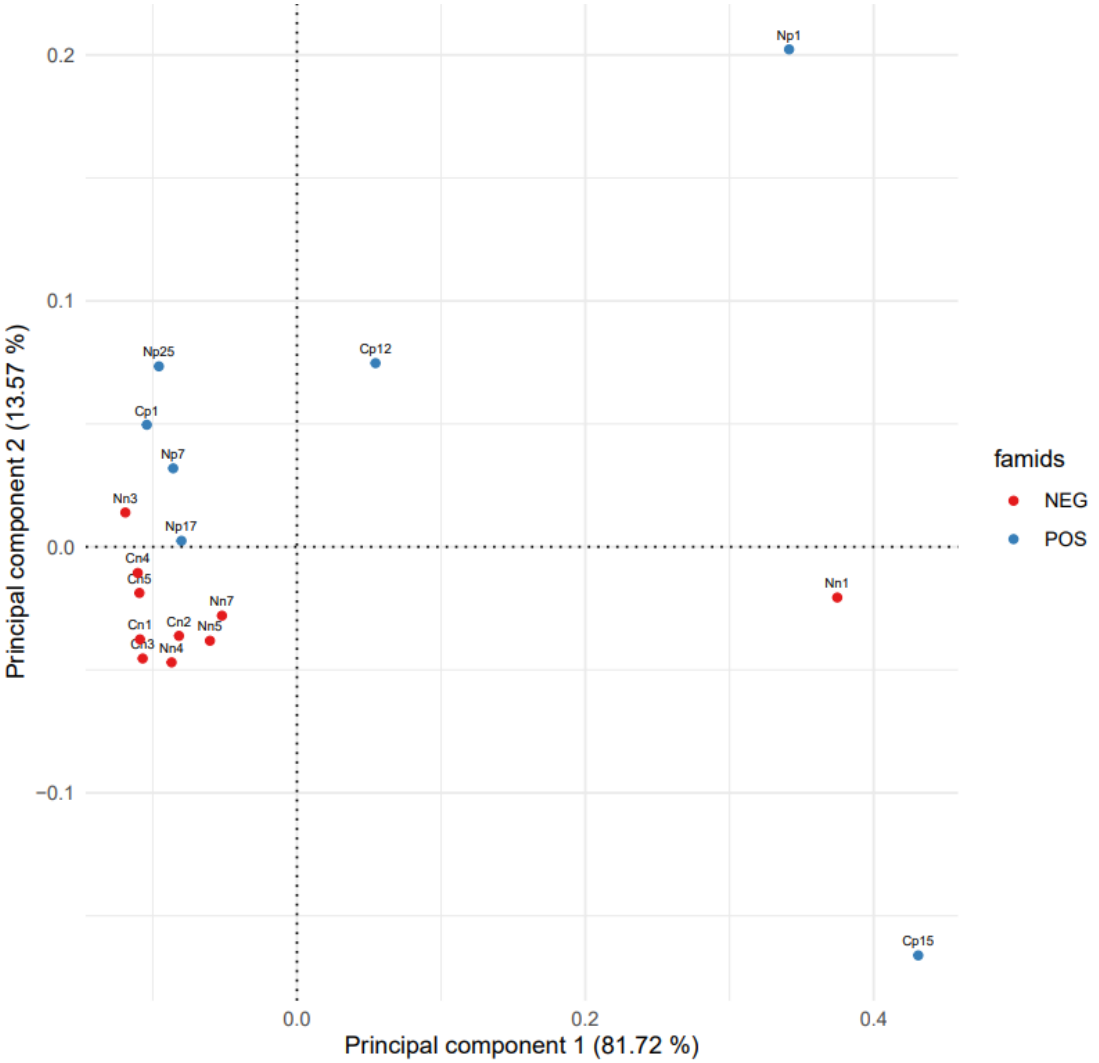


Figure 8. *Principles component analyses of the 17 post-filter Cryptosporidium individuals. PC A1 (81.72%) and PCA 2 (13.57%) shows 1 distinct cluster, with 4 outliers: individuals Np1, Cp12, Cp15 and Nn1.*

The outlier individuals, all stool samples, were likely due to genotyping errors because of a difference in quality. Np1 had the lowest proportion of missing SNPs at 0.197%, as well as comprising the upper quartile of DNA yield (29,255ng), 260/280 (1.89) and 260/230 (2.02). Cp15 also

had a low proportion of missing SNPs at 0.2556%, but conversely had also had a low DNA yield of 890ng- indicating that despite a low DNA yield the samples may contain a large quantity of cattle DNA. Sample Nn1 had a higher proportion of missingness (0.442%) as well as a minimal DNA yield of 930ng and poor quality, with 260/280 at 1.57 and 260/230 at 0.54, indicating poor sample quality and poor adhesion to the SNP, indicating contamination. Cp12 had an unusually high yield of 15,090ng, and favourable quality ratios of 260/280 at 2.08 and 260/230 at 1.98- considering this sample also had a high rate of missingness at 0.382%, this may indicate a large quantity of microbial DNA present, affecting SNP Chip efficiency.

3.3 (a) FST Analysis.

With the data filtered, we performed an FST analysis to identify regions-of-interest. A total of 34,455 SNPs with valid FST estimates were included to identify genomic regions of high divergence between POS and NEG individuals, with a mean FST estimate of 0.050428, and a weighted Fst estimate of 0.0560627 (*Figure 9).*

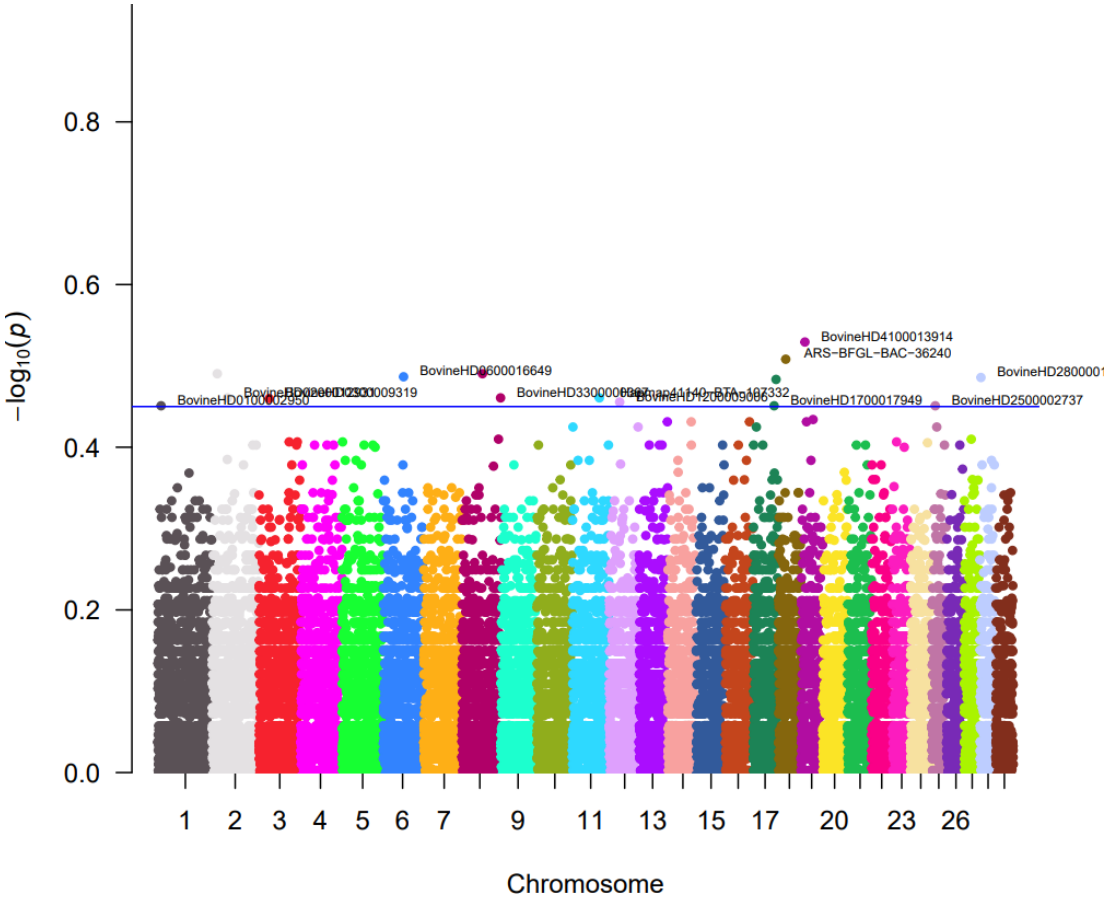


Figure 9*. Pairwise FST Manhattan plot of POS vs NEG cattle, giving FST estimates per SNP per position per chromosome (chromosomes differentially coloured). The suggestive line is plotted at FST 0.45.*

Regions of Interest (RoI), denoted by SNP per position per chromosome, were denoted as having ≥0.45FST, returning 15 regions for investigation. The coordinates of these regions are given in *Appendix Table 2*. For each position, genes located in 50,000bp windows up/downstream of the RoI were identified as Genes of Interest (GoI), with 23 unique protein coding GoI (*Table 5*), 22 of which were characterised for further analysis.

3.3(b) Gene Ontology Analysis.

The ENSEMBL Gene IDs were analysed according to the Panther Ontology functional classification system (*Table 5*) as well as mining through relevant literature. These genes may represent putative candidates for either susceptibility and/or resistance to *Cryptosporidium* and *Cryptosporidium*-related infection/illness, presenting areas for further research.

***Table 5**. List of 23 genes with corresponding gene name, when available (otherwise are 'Uncharacterised'). The resulting characterised proteins presented 22 genes for further analysis.*

ENSEMBL Gene ID	Gene Name	Protein PANTHER class (when available)	Start position (bp)	End position (bp)
ENSBTAG00000049117	(Uncharacterised)		11:78441028	11:78445794
ENSBTAG00000011155	PUM2	RNA Metabolism protein	11:78486240	11:78601472
ENSBTAG00000031824	RBM19	RNA Splicing Factor	17:60618709	17:60739666
ENSBTAG00000010368	TPST2	Transferase	17:66190152	17:66244471
ENSBTAG00000019280	CRYBB1	Structural Protein	17:66251943	17:66289558
ENSBTAG00000019282	CRYBA4	Structural Protein	17:66252289	17:66275059
ENSBTAG00000002287	CHD9	chromatin/chromatin-binding, or -regulatory protein	18:21555470	18:21782071
ENSBTAG00000009968	TBX4	Rel homology transcription factor	19:11607447	19:11633858
ENSBTAG00000014278	TBX2	Rel homology transcription factor	19:11683158	19:11692577
ENSBTAG00000006907	NEB		2:44434401	2:44651653
ENSBTAG00000000937	ITPRID2		2:4654230	2:14751034
ENSBTAG00000005453	FGD4	GTPase activity	2:77603357	2:77833309
ENSBTAG00000025021	FMN2		28:2953439	28:3309533
ENSBTAG00000018881	SYT6	Membrane trafficking regulatory protein	3:29231274	3:29297466
ENSBTAG00000017809	PDS5A	Chromatin/chromatin-binding, or -regulatory protein	6:58887610	6:59020226
ENSBTAG00000011420	CA9	Dehydratase	8:59849175	8:59855369
ENSBTAG00000011424	TPM2	Actin binding motor protein	8:59856503	8:59864645
ENSBTAG00000025868	TLN1		8:59869867	8:59901292
ENSBTAG00000011429	CREB3		8:59901572	8:59905938
ENSBTAG00000011431	GBA2	Glucosidase	8:59901585	8:59921540
ENSBTAG00000011433	RGP1		8:59917058	8:59922869
ENSBTAG00000038777	MSMP	Intercellular signal molecule	8:59921093	8:59922048
ENSBTAG00000050718	TMEM18		8:112472379	8:112479359

The 22 characterised genes were grouped according to their Panther GO-Slim biological process category (*Figure 10*). These genes spanned 47 total process hits, with cellular process (GO:0009987)

having the highest number of associated genes (13 genes), followed by metabolic process (GO:0008152) (7 genes).

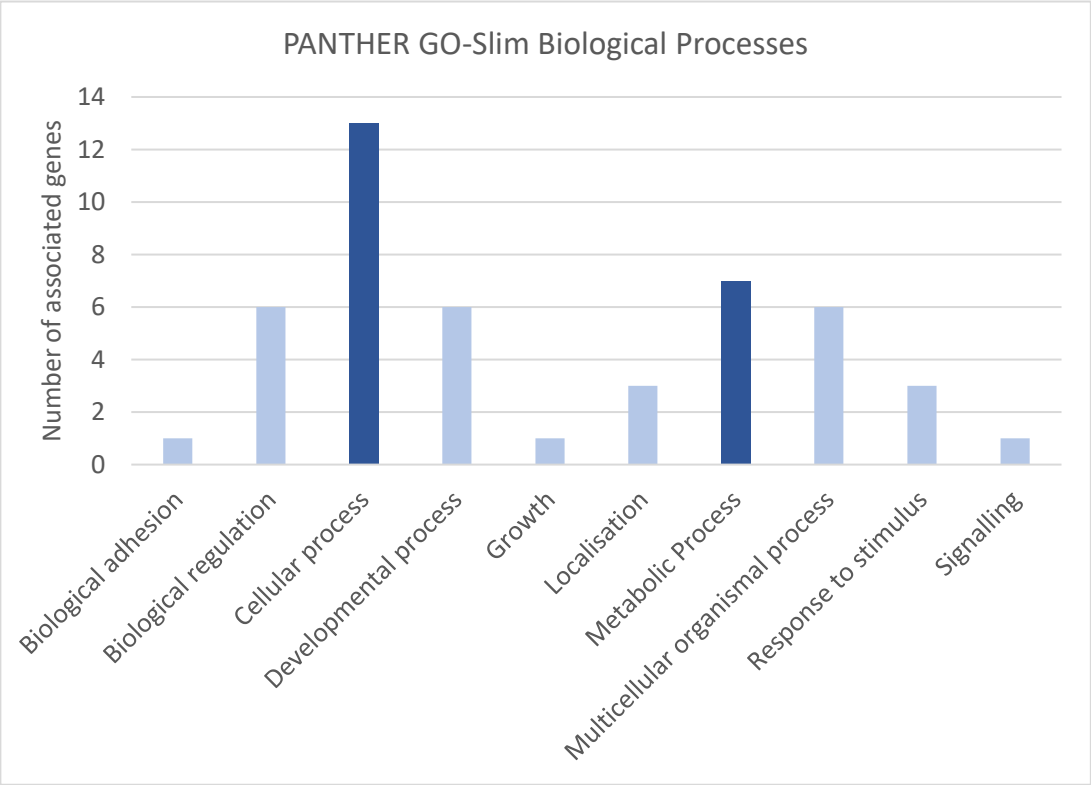


Figure 10. *Panther GO_Slim biological processes for 22 characterised protein-coding genes putatively associated with Cryptosporidium, with processes with highest number of associated genes highlighted.*

3.4 Asian Cattle Genetic Diversity.

To identify the population structure of the Asian breeds, and the relatedness of the Thailand cattle, the samples were clustered according to breed, each of which originated from different regions in Asia (*Figure* 2). The raw dataset comprised 6 breeds, a total of 46 individuals, and a total of 27,371 SNPs. The DRYAD dataset, filtered according to QC1, resulted in 26,069 SNPs and all 41 individuals passing quality control. The Thailand cattle, filtered according to QC2, resulted in 26,402 SNPs and 3 individuals passing quality control. Merging this dataset and removing duplicate SNP positions resulted in 44 Asian cattle individuals and 25,973 SNPs remaining for analysis.

3.4 (a) Population Structure of Asian cattle.

After filtering using the whole dataset, 22,407 SNPs remained for analysis. We calculated FST values across all breeds, with a mean FST of 0.276807, and a weighted FST of 0.300351. We then compared each population to our Thailand samples by estimating pairwise FST values (*Table 6*). Our results indicate that the Mongolian (Northernmost Asian breed of this dataset) had the lowest FST score and hence greatest genetic similarity with the Thailand population, while the Brebes (Southernmost breed of this dataset) had the highest FST score and hence lowest genetic similarity with the Thailand population.

Populations	Mean FST score	Weighted FST score
THA_MON	0.288579	0.439846
THA_HAN	0.306861	0.455351
THA_WAG	0.347031	0.509239
THA_MAD	0.38328	0.601924
THA_BRE	0.390813	0.607903

Table 6. *Measuring population differences between the Thailand cattle and other cattle breeds via pairwise FST analysis, given in ascending order by FST score.*

To further study population structure, we used a PCA test. The PCA formed 3 distinct clusters (*Figure 11)*, with the Southernmost breeds (MAD and BRE), both of Indonesian origin, forming a cluster; the Northernmost breeds forming a cluster (WAG, HAN and MON) and THA cattle forming a separate cluster. Within these clusters, BRE and MAD formed an overlapping cluster, indicating low intra-breed genetic variability. PC1 explains 32.74% of the variance, and it separates the North and South breeds, while PC2 explains 18.52% of the variance and clearly separates Thailand samples from the other breeds.

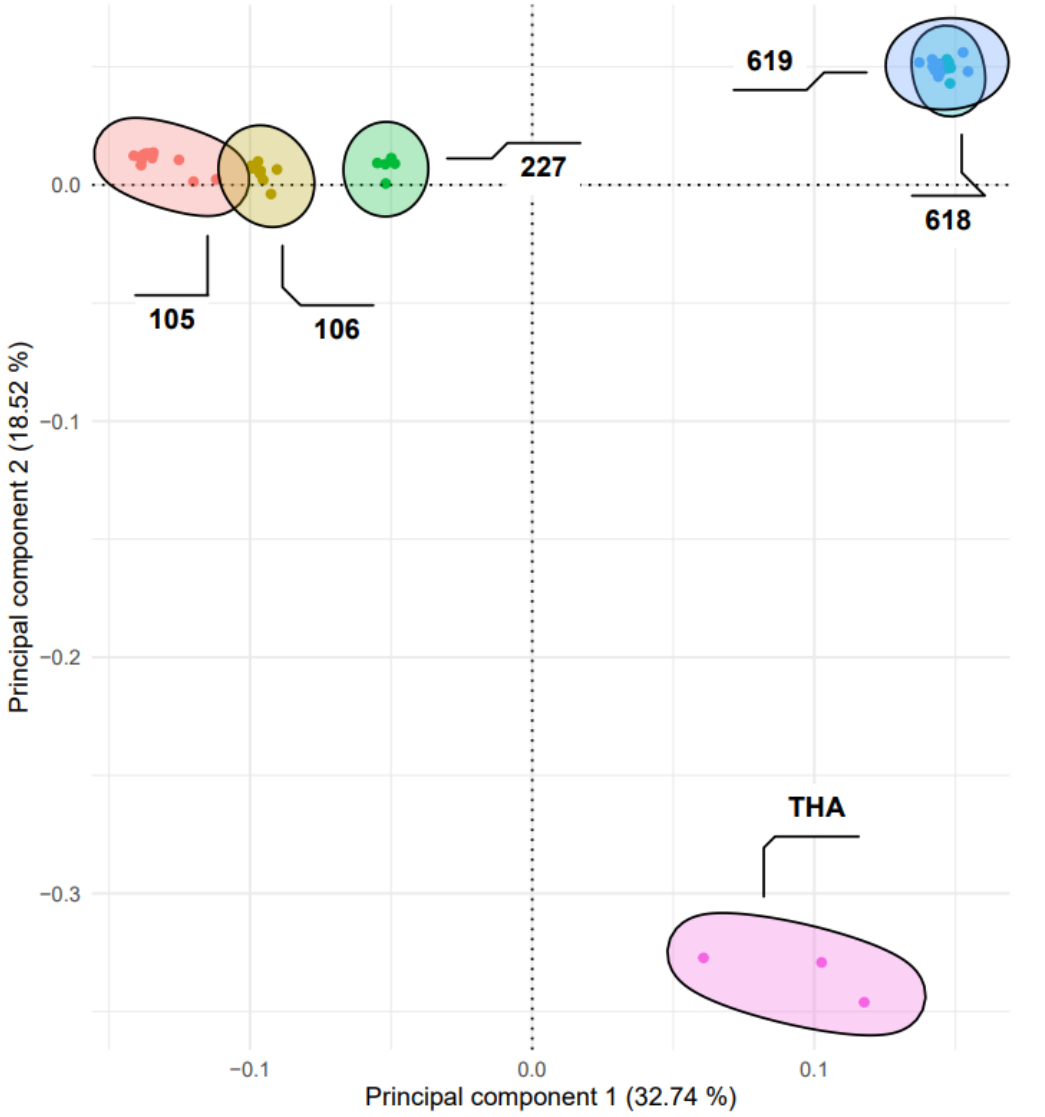


Figure 11. *Principle component analysis of 6 Asian Cattle Breeds. PCA 1 (18.52%) and PCA 2 (32.74%) shows 3 distinct clusters, with BRE (618) and MAD (619) breeds overlapping.*

We then ran STRUCTURE to further estimate the patterns of genetic variation among the subpopulations (*Figure 13*). The Madura and Brebes breeds were genetically identical across all K assumptions and had minor admixture with the Mongolian (K=3, K=6, K=8) and Hanwoo (K=3, K=6). For K=3, K=6 and K=8, the Thailand cattle also appear as a genetically distinct subpopulation from all other breeds. For K=3 and K=6 the Thailand cattle demonstrated some admixture with other Northern Cattle subpopulations, specifically the Hanwoo and Mongolian breeds.

We used the STRUCTURE Harvester to estimate the probability for best K-value parameter (*Figure 12*). STRUCTURE Harvester results suggested that the greatest likelihood is 3 subpopulations comprising the total population. The highest Delta (K) was 10,636.96 for K=3 and the lowest Delta (K) was 0.44 for K=4. For L(K), all K assumptions greater than K=3 had lower accuracy (mean of est.LB prob of data = <0) and greater variance, indicating fine-scale substructure that Delta(K) is not sensitive to.

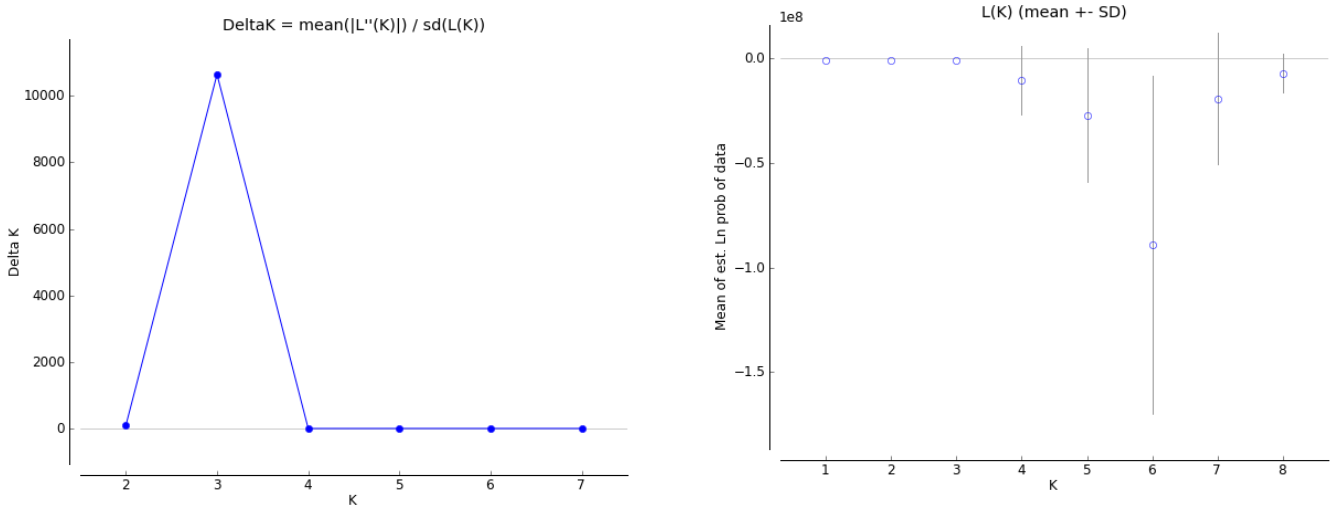


Figure 12. Results of STRUCTURE Harvester.

(A) *Ad hoc quantity (ΔK): Calculated based on the second order rate of change of the likelihood (ΔK) (Evanno et al. 2005). The ΔK shows a clear peak at the true value of K. ΔK = m([L''K])/s[L(K)]*

(B) *Use of L(K): When K is approaching a true value, L(K) plateaus (or continues increasing slightly) (Rosenberg et al. 2001); mean of est.LB prob of data decreases after K=3.*

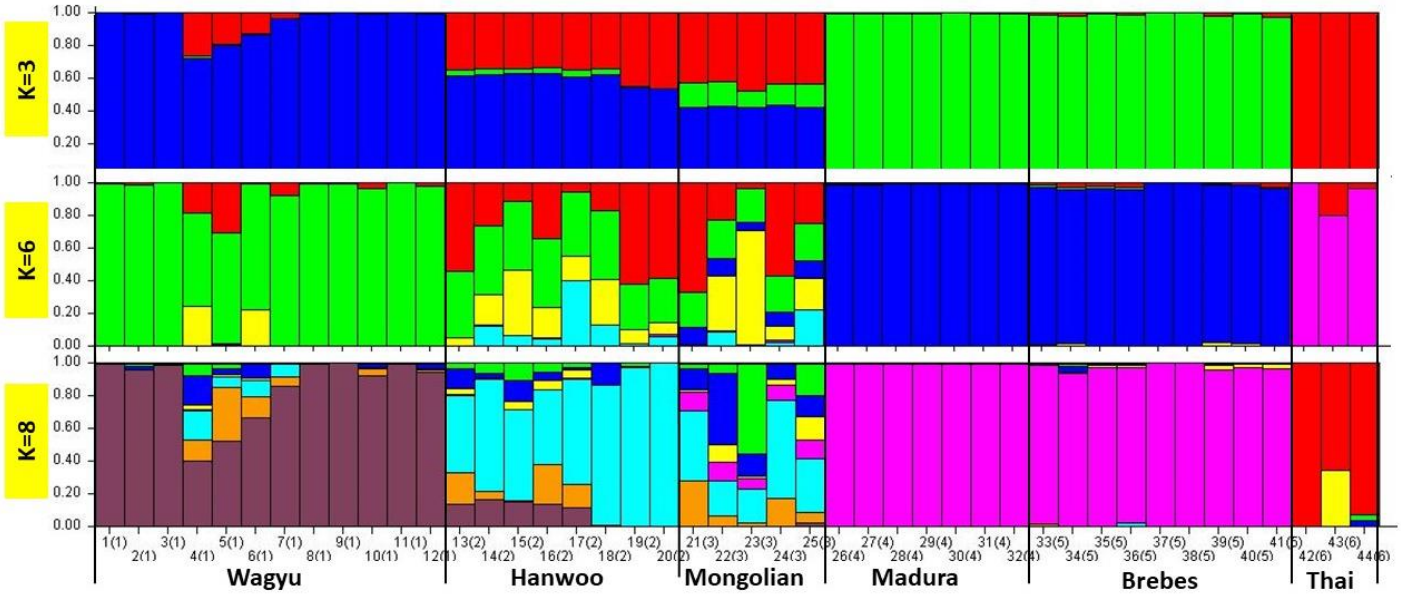


Figure 13. *STRUCTURE results for the 6 Asian cattle breeds, for assumptions of K=3, K=6 and K=8. According to the results from STRUCTURE Harvester, K=3 presents the most likely number of subpopulations.*

3.4 (b) Evolutionary relationships between populations.

Finally, to establish the relationships between populations we used TreeMix analysis, the maximum likelihood tree of the 6 Asian subpopulations (+1 outgroup subpopulation, Holstein-Friesian, *FRA*) was inferred (*Figure 14*). Migration events were added to the tree sequentially from 0 to 5 migrations (all migrations are shown in *Appendix Figure 1*). For all migrations, the Thailand cattle consistently formed a separate branch with no migration events. Brebes and Madura consistently formed sister branches. Wagyu consistently demonstrates introgression with the *taurine* outgroup. The North Asian cattle breeds (excluding Thailand) demonstrated variability in tree position depending on migration edges. The most likely number of migration edges is 3, shown in *Figure 13*, which was estimated using OptM.

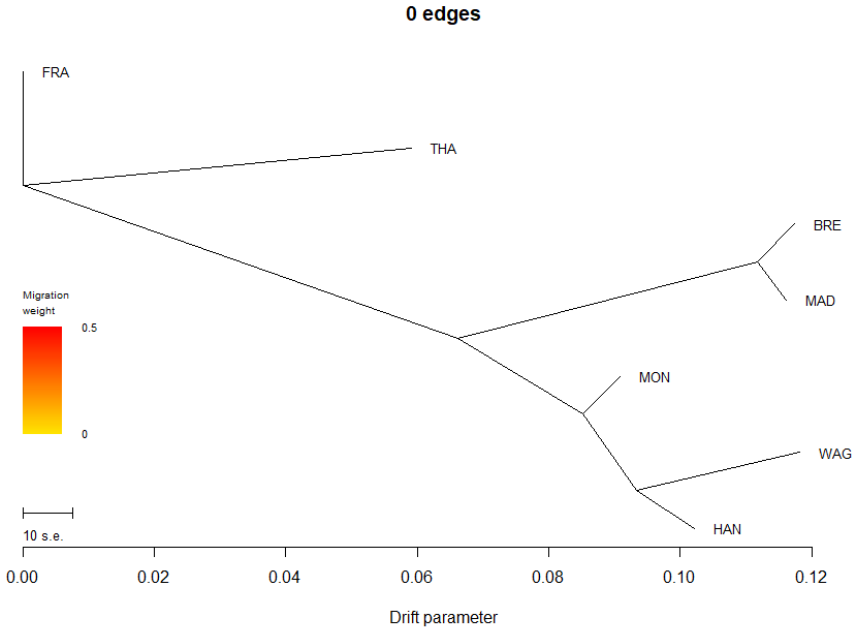


Figure 14 (A). *Shows the maximum likelihood tree of the 6 Asian subpopulations, rooted to the Hostein-Friesian outgroup.*

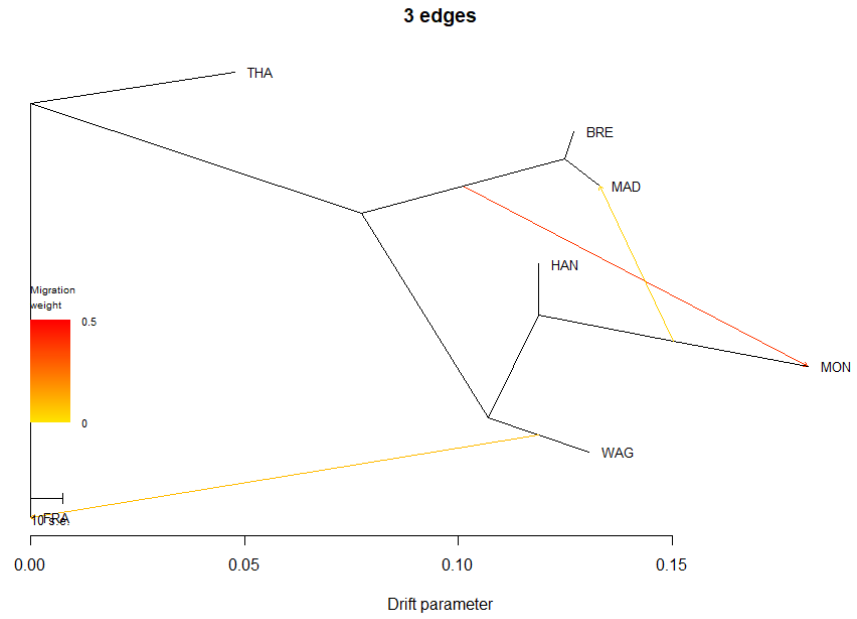


Figure 14 (B). *The most likely number and structure of migration edges between the 6 Asian subpopulations and 1* Bos taurus *outgroup population.*

Figure 14. *Population graphs obtained with TreeMix. Holstein-Friesian of Bos taurus ancestry ('FRA') was denoted as the root (outgroup). For the 6 Asian subpopulations abbreviations see Table 1.*

4.0 Discussion.

4.1 Sample suitability for SNPChip genotyping.

DNA was extracted from three sample types: stool, blood and nasal swabs. To assess DNA quality, we used a Nanodrop. Readings of the 260/280 ratio are optimally ~1.8 (a higher or lower ratio indicates contamination with protein, phenol or other contaminants). The 260/230 ratio is optimally ~2.0-2.2, and is used to indicate the presence of unwanted organic compounds; values outside of this range may indicate contamination.

Quality assessment revealed that stool samples had the poorer quality DNA. Whilst the 260/280 ratio was near optimal, the low 260/230 ratios indicated contamination (*Figure 5*).

Stool DNA is typically considered an advantageous sample type, as it is a non-invasive and usually easily obtained. However, as has been confirmed in previous research, DNA from stool may be degraded due to environmental exposure and contamination by nontarget, that is, microbial DNA[47,48]. Whilst the PCR results confirmed that host (cattle) DNA is indeed present in the sample (*Figure 6*), a qPCR should be performed on future DNA extractions of stool samples, to confirm that the required yield of host DNA is present. Previous studies have found that often less than 5% of the total DNA extracted from stool contains host DNA, with most DNA coming from pathogens, parasites and commensal bacteria[49]. Considering this, although we extracted an average of 5220 ng DNA from stool samples, host DNA yield was probably low. This, combined with low DNA quality, might explain the high %Missing Call Rate for stool samples and the need to use less-than-optimal quality SNP calling filters in this research, which may have led to spurious results due to undetected genotyping errors.

If DNA extracted from stool samples must be used, it is recommended to conduct a pilot study to quantify genotyping error rates for each population to be studied, as well as season; one study on wild ungulates detected higher genotyping error rates for spring stool than for winter stool[50].

Blood samples had the lowest average DNA yield. Whilst the contamination of proteins 260/280 ratio was optimal, the contamination of other, unwanted organic compounds (260/230) ratio was also present. Considering that blood is typically associated with high-molecular weight DNA suitable for genotyping, the poorer quality 260/230 ratio was unexpected[51]. Prior research has suggested that low 260:230 ratios may be due to EDTA contamination from the collection tubes, which absorbs strongly at 230 nm[52]. Despite these contaminants, it is unlikely that these substantially affected downstream analyses, because blood samples had a substantially lower %Missing call rate than the other samples used in this research, indicating the high suitability of blood samples for SNP Chip genotyping.

Nasal swabs produced the highest quality DNA of all sample types used in this research. Our results mirror a previous comparison of DNA collection methods and sample types, also in cattle (and yaks), where the authors found nasal swabs to produce higher quality DNA[52]. The nasal swab extractions also produced a high yield of DNA. Despite this, DNA extracted from nasal swabs had unexpectedly high and varying Missing Call Rates, indicating poor suitability for SNP-Chip genotyping. One possible explanation is because all sample types were genotyped on an array mapped to a *Bos taurus* genome assembly. Hence, the SNP-Chip used has a strong ascertainment bias towards *Bos taurus* breeds. The nasal swab samples that originated from Thailand are of *Bos indicus* type cattle, which are genetically distinct from *Bos taurus* breeds. Novel commercial SNP Chips for *Bos* indicus

cattle are available; previous research compared the performance of *Bos taurus* vs *Bos indicus* SNP chips on indicine cattle from India and concluded that SNP panels with more SNPs polymorphic to *Bos indicine* cattle (i.e the GeneSeek 75K *indicus* chip) represents the better choice for *Bos indicus* breeds[53]. This hence explains the higher %Missing Call Rate observed for the Nasal Swab (*Bos indicus*) samples. Future studies should consider the availability of better suited SNP Chips for their breed of animal.

4.2 European *Cryptosporidium* Association Analysis.

Although this analysis included two geographically separate cattle populations (from the Netherlands and from Cyprus), the genetic differentiation between the cattle was low, as indicated by the low FST values (see 3.3). Holstein-Friesian cattle breeding is systemically selected and controlled. The populations belong to the same breed and originate from the same recent ancestral population; this may explain the low FST value, since it is likely therefore that many SNPs are commonly fixed across the populations. Greater pedigree information would further elucidate these results.

4.2 (a) Gene Ontology Analysis.

A total of 15 regions with FST >0.45 were identified using 34,455 SNPs of 17 individuals. Using 50,000bp windows up/downstream of the region, 23 unique protein coding genes were returned, 22 of which were characterised. A GO Enrichment analysis using PANTHER classification system revealed cellular process (GO:0009987) to have the highest number of associated genes (13 genes). This term is associated with cellular processes involved in cell growth and/or maintenance, and cell physiology. Notably, *Cryptosporidium* invades host epithelial cells by altering the hosts cellular processes; it induces rearrangement of the host cell cytoskeleton. Specifically, it integrates host cell actin and α-actinin, forming host-parasite interface via a junctional complex[54]. Although more in-depth research is required, for example with larger scale data, this indicates that resistance to *Cryptosporidium* infection possibly has a genetic basis.

Metabolic processes (GO:0008152) also had a significant number of associated genes (7 genes). This term includes genes associated with chemical reactions and pathways by which living organisms transform chemical substances. Research by *Vélez et al (2021)*[55] used in-vitro gene-expression based evidence in humans to reveal metabolic signatures of *Cryptosporidium parvum* infection. The research found a down-regulation of host-parasite homologous genes which are associated with host glycolysis/gluconeogenesis pathways, whilst host-exclusive genes were upregulated, suggesting parasite-derived direct competition for metabolic processes. Considering the results of our research, it may be possible that there are similar metabolic signatures present in *Cryptosporidium* infection in cattle, although this would require further research.

Of the 22 characterised protein-coding genes in or near RoI, 5 are putative candidates for association with *Cryptosporidium* infection based on analysis of previous research findings for these genes:

FGD4 (RhoGEF and PH domain-containing protein 4): FGD4 encodes a protein involved in the regulation of cytoskeleton and cell shape. This gene is also involved in the activation of Cdc42 via guanosine triphosphate exchange; Cdc42 has previously been demonstrated to be constitutively activated *in vitro* in *Cryptosporidium* infected cells, recruited to the host-parasite interface[56]. Depletion of Cdc42 mRNA by short interfering RNA-mediated gene silencing was also found to inhibit *Cryptosporidium* infection. Hence, this finding presents a potential gene target for modulating successful *Cryptosporidium* infection but has yet to be applied to *in vivo* experiments.

FMN2 (Formin 2): This gene encodes a protein which is associated with essential roles in organization of the actin cytoskeleton and in cell polarity. Whilst no prior research has found a link between this gene/protein and *Cryptosporidium* infection, considering the importance of host cytoskeletal rearrangement to infection by *Cryptosporidium*, this would benefit from further investigation.

CA9 (Carbonic anhydrase 9): Carbonic anhydrases (CA) are a large family of zinc metalloenzyme that catalyse the reversible hydration of carbon dioxide.
Specifically, CA9 is mainly expressed in the gastrointestinal tract where it facilitates acid secretion by gastric parietal cells. It is worth noting that, once ingested by the host, *Cryptosporidium* oocysts excyst in the gastrointestinal tract; due to stomach acid exposure, *Cryptosporidium* releases infective sporozoites[57]. Understanding how CA9 is regulated during *Cryptosporidium* infection would further clarify its role.

The promoter region of the CA9 gene contains an HRE (hypoxia responsive element) where HIF-1 can bind, enabling hypoxic conditions to increase. Hypoxic conditions have been associated with host response to *Cryptosporidium* infection previously[58, 59]. Downregulation of a different carbonic anhydrase isoform, CA4, has been found related to fatal infectious diarrhoea in mice, due to ensuing defects in intestinal chloride absorption[60]. This finding validates the importance of acid regulation in *Cryptosporidium* infection, as well as provides the basis for further research into the interaction between the numerous CA isoforms and *Cryptosporidium* infection.

TPM2 (Tropomyosin 2): Tropomyosin 2 encodes beta-tropomyosin, a member of the actin filament binding protein family, and is mainly expressed in intestinal smooth muscle fibres; all isoforms play an integral role in actin filament dynamics. *O'Hara, S et al (2006)*[61] investigated the roles of various tropomyosin isoforms in *Cryptosporidium* infection using isoform-specific monoclonal antibodies to detect tropomyosin expression in cultured human HCT-8 cells. The result found an accumulation of Tropomyosin isoform *TM5* at the infection sites of these host cells during *Cryptosporidium* infection, but the study was limited in that these cells did not express *TM2*. Further research should be attained to determine which tropomyosin isoforms are expressed in cow epithelial cells, and further to this, if (and then which) tropomyosin isoform(s) are expressed during *Cryptosporidium* infection.

TLN1 (Talin 1): TLN1 is associated with cytoskeletal proteins that are concentrated in areas of cell-cell adhesions. The *N-terminus* of this protein contains elements for localization to cell-extracellular matrix junctions, and the *C-terminus* contains binding sites for proteins such as actins[62]. Talin has previously been identified in the host-cell actin accumulations associated with enteropathogenic infections via focal adhesion-association actin-binding molecules[63,64,65]. Therefore, further investigations into the expression of TLN1 would benefit this understanding (i.e TLN1 may be downregulated during *Cryptosporidium* infection).

The discussion on gene ontology and individual genes offers a putative assessment of *candidate* genes only, which are associated with *Cryptosporidium* infection; it cannot be definitively ascertained from this preliminary investigation whether these genes are up- or down- regulated, or whether associated with susceptibility or resistance. However, the results of this research provide further groundwork suggesting that resistance/susceptibility to *Cryptosporidium* infection in cattle does have a genetic basis, and justifies further research into this area.
Methods to find out the expression of activated genes (i.e detection of relevant mRNA *in vitro,* or an RNA-seq *in-vivo* experiment with intestines of infected vs noninfected cattle) could benefit and extend this research (see *Cekan et al (2004)*[66] for further suggestions on methodology).

Furthermore, yet greater validation of these results and discussions could be attained with larger scale data.

4.3 Asian Cattle Breed Genetic Diversity.

The aim of this research was to investigate and understand the genetic diversity and evolutionary of the Thailand cattle in the context of other Asian cattle breeds, and in doing so, consider relevant genetic management strategies, and whether the Thailand cattle constitutes a genetic resource.

We extracted the DNA from nasal swabs of 5 Thailand zebu-type cattle. After SNP calling filtering, 3 Thailand Cattle individuals remained for analysis with 26,403 SNPs. To assess the Thailand cattle in a wider geographical context, we compared the SNP data to priorly published data of 5 other Asian breeds, a DRYAD dataset of 26,069 SNPs and 41 individuals (see *Table 1*), comprising the Hanwoo breed (Korea), Mongolian bred (Mongolia), Brebes bred (Java, Indonesia) and Madura (Madura, Indonesia).

The results found the Thailand cattle to be genetically distinct among the 6 Asian breeds investigated.

According to K=3, the most likely number of populations suggested by STRUCTURE Harvester, Thailand cattle form a genetically distinct population (*Figure 13*). Thailand cattle are also genetically distinct for K=6 and K=8, albeit demonstrating some admixture with other Northern Cattle subpopulations, specifically the Hanwoo and Mongolian breeds.

This agrees with our TreeMix analysis, where for all 0-6 migration edges in the TreeMix analysis, Thailand cattle formed their own branch separate from both the *Bos indicus* breed groups and the *Bos taurus* outgroup (*Appendix Figure 1*). For the most likely number of migration edges (3), Thailand cattle demonstrated no migration with other breeds. Only at 5 migration edges is there is some admixture between an ancestor of the Mongolian breed and the Thailand population; this clarifies the lower FST scores between Mongolian and Thailand. Such results may suggest that the Thailand cattle genotyped present a unique breed, which require genetic management for their conservation.

Ariyaraphong N, et al (2021)[19] used ancient DNA and large-scale MT D-Loop sequencing in Thailand dairy cattle, and found evidence indicating that Thailand cattle breeds originated from China and were introduced to central Thailand between 3550 and 1700 years ago. Of the 6 Asian breeds, Mongolian breeds are geographically closest to China (as well as Thailand). The Mongolian breed is native to Inner Mongolia, as well as to northern China[67,68]. Considering this, it is very likely that cross-breeding between the Thailand cattle and Mongolian breed has occurred. Other research has also found that a variety of Thailand cattle, the Kho-Chon variety (mostly found in Southern Thailand), appears to be a genetically distinct breed with minimal admixture with other Asian breeds[69]. Unfortunately, out samples were from an unknown breed (or variety) origin, impeding further conclusions- but it may be likely that our sampled Thailand cattle were of Kho-Chon variety, considering their unique genetic structure compared to the other Asian breeds included in this dataset. This research hence justifies investigating the Thailand cattle varieties as independent breeds. Future studies should use a wide range of Thailand cattle to determine where breed status might be due.

Furthermore, previous research has shown that Thailand cattle demonstrate at least some admixture with purebred Holstein-Friesian, due to local cross-breeding programmes in the past 60 years[17,18, 69]. This was not observed in our results. The Thailand cattle we genotyped may be a unique set of individuals, or it may highlight the need to include a greater sampling of Thailand individuals, as well as more generally the need to improve pedigree records for cattle in Thailand.

The Southernmost breeds used in this dataset, Brebes and Madura, are geographically isolated from the Thailand cattle, as well as the other Asian breeds used in this dataset; admixture between Madura/Brebes and other genotyped breeds is minimal as observed the STRUCTURE plots (Brebes shows minimal admixture with other breeds) (*Figure 13*). Along with the geographic isolation, starting early in the 20th century Madura/Brebes breed have been maintained pure on the islands of Madura and Brebes by closing the islands from other breeds[70]. Our results clearly demonstrate the Madura/Brebes breed to be genetically indistinct from each other and highly genetically distinct from the other Asian breeds included in this analysis, as demonstrated in the PCA (populations overlapping) (*Figure 11*) and STRUCTURE analysis (*Figure 13*). This is also in agreement with previous research[70,71].

For all 0-6 migration edges in TreeMix, the Brebes and Madura breeds consistently form sister branches (*Appendix Figure 1*). For the most likely number of migration edges (3), as well as for all migration edges, the common ancestor for the Brebes and Madura breed demonstrates admixture with the present-day Mongolian breed, with a high migration weight. The opposite is also observed, admixture from the ancestor of the Mongolian to the Madura breed. This relationship has not been demonstrated by any previous research, justifying greater research into the relationship between the cattle of Indonesia and Mongolia- for future research, greater sampling is recommended to assuredly elucidate on this question.

The North Asian cattle breeds (Hanwoo, Wagyu and Mongolian) demonstrated variability in tree position depending on number of migration edges, but consistently formed a group. This is also demonstrated in the PCA analysis (these three breeds form a loose cluster). The STRUCTURE results reveal all the Northern breeds to be genetically diverse with varied intra- and inter- admixture the Northern breeds. Previous research by *Deker J, et al (2014)*[71] reveals evidence for European admixture *(Bos taurus)* for these breeds, with Wagyu specifically demonstrating ~0.188% of their genome originating from Northwestern European ancestry. For the most likely number of migration edges (3), the ancestor of the Wagyu breed shares admixture with the outgroup, modern-day French *(Bos taurus)* cattle; this is also observed for 4 and 5 migration edges. For 2 migration edges, there is admixture between the ancestor for all three Northern cattle breeds and the *Bos taurus* outgroup. In line with previous research, our findings further validate *Bos taurus* admixture with Northern Asian cattle breeds[71].

For these investigations, numerous plots of STRUCTURE are presented to interpret the population structure of the 6 Asian breeds. Whilst STRUCTURE Harvester determined there to be 3 populations (K=3) present within the dataset – the same outcome as populations observed in the PCA- the STRUCTURE programme is not without its limitations. STRUCTURE Harvester has been demonstrated to miss fine-scale structure, which is especially true with inappropriate sampling[72]. Hence, the results and discussions here would benefit from larger scale data, and where this is not possible, information on individual breed pedigree would also aid understanding the breeds relationships.

Conclusion.

To continue exploiting cattle for their nutritional benefit on a global scale, modern agricultural practices must maximise yield output on land that is already available, to limit the burdens of climate change. To discuss methods for this approach both in terms of large-scale commercialised agriculture and in terms of small-scale, localised agricultural practices, a fast and affordable method of obtaining genetic information was used (SNP Chip genotyping).

SNP Chip genotyping data offers a fast and affordable method to gather genetic information to answer genetic diversity and population genetics research. However, our research confirms that for this to be effectively utilised, high-quality DNA extracted from blood and/or nasal swabs is ideal. This research was also limited by small sample collections, and a lack of available metadata, especially pertaining to pedigree and breed/variety information-- this is a known problem for lack of appropriate record-keeping in local Thailand farming practices, presenting an important area for improvement, for example, via educational outreach. Understanding these limitations, these results lay groundwork for improvement with further research.

The local adaptations of Zebu-type cattle in Asia present a valuable resource for genetic resistance under climate change, but for this to be effectively utilised, greater understanding of the genetic structure of localised breeds (including the Thailand cattle) must be achieved. The population structure underlying several Asian breeds was hence elucidated and supplemented previous research, identifying high genetic diversity for Indonesian breeds and greater admixture for the Northern Asian breeds. Interestingly, Thailand cattle were found to be genetically distinct. This is not unheard of for the *Kho-Chon* Thailand variety and agrees with previous research efforts which identify Thailand cattle as unique among other *Indicine* breeds.

In addition, improvements to modern industrial dairy practices were also investigated, in terms of improving global food security. This research demonstrated that host genetic resistance to parasitic infection could provide a useful tool in efficiently increasing yield output for dairy farms. Putative genes associated with *Cryptosporidium* were identified, primarily involved in cellular processes and metabolic processes, laying the groundwork for further investigations: FMN2, TPM2 and TLN1 (novel), and CA9 and FGD4 (previously found to be directly/indirectly associated with *Cryptosporidium* infection).

In conclusion, the use of SNP Chips for genetic investigations present a time-conscious and economically viable means of meeting important sustainable goals, including improving global food security and maintenance of genetic resources, both important factors under climate change conditions. However, as our research demonstrates, for genetic resources to be used to their full potential, large-scale, high-quality samples must be attained. Further, when appropriate SNP Chip data should be supplemented with good pedigree-record keeping to better understand the relations of the individuals being investigated.

References

1. FAO, IFAD, UNICEF, WFP and WHO. 2021. 'The State of Food Security and Nutrition in the World 2021. Transforming food systems for food security, improved nutrition and affordable healthy diets for al'l. Rome, FAO. Available at: https://doi.org/10.4060/cb4474en

2. Vågsholm, Ivar & Arzoomand, Naser & Boqvist, Sofia. (2020). Food Security, Safety, and Sustainability—Getting the Trade-Offs Right. Frontiers in Sustainable Food Systems. Available at: https://doi.org/10.3389/fsufs.2020.00016

3. Ajmone-Marsan, P., Garcia, J.F. and Lenstra, J.A. (2010) 'On the origin of cattle: How aurochs became cattle and colonized the world', Evolutionary Anthropology: Issues, News, and Reviews, 19(4), pp. 148–157. Available at: https://doi.org/10.1002/evan.20267.

4. Felius, M. et al. (2011) 'On the Breeds of Cattle—Historic and Current Classifications', Diversity, 3(4), pp. 660–692. Available at: https://doi.org/10.3390/d3040660.

5. Groeneveld, L.F. et al. (2010) 'Genetic diversity in farm animals--a review', Animal Genetics, 41 Suppl 1, pp. 6–31. Available at: https://doi.org/10.1111/j.1365-2052.2010.02038.x.

6. Bradley, D.G. et al. (1996) 'Mitochondrial diversity and the origins of African and European cattle.', Proceedings of the National Academy of Sciences of the United States of America, 93(10), pp. 5131–5135.

7. Diamond, J. (2002) 'Evolution, consequences and future of plant and animal domestication', Nature, 418(6898), pp. 700–707. Available at: https://doi.org/10.1038/nature01019.

8. Pitt, D. et al. (2019) 'Domestication of cattle: Two or three events?', Evolutionary Applications, 12(1), pp. 123–136. Available at: https://doi.org/10.1111/eva.12674.

9. Uberoi, E. et al (2021) 'UK Dairy Industry Statistics', House of Commons Library UK, Available at: https://commonslibrary.parliament.uk/research-briefings/sn02721/

10. Taberlet, P. et al. (2011) 'Conservation genetics of cattle, sheep, and goats', Comptes Rendus Biologies, 334(3), pp. 247–254. Available at: https://doi.org/10.1016/j.crvi.2010.12.007.

11. Berthouly, C et al. "Revealing fine scale subpopulation structure in the Vietnamese H'Mong cattle breed for conservation purposes." BMC genetics vol. 11 45. 7 Jun. 2010, https://doi:10.1186/1471-2156-11-45.

12. Rischkowsky, B., Pilling, D. and Nations, F. and A.O. of the U. (2007) The State of the World's Animal Genetic Resources for Food and Agriculture. Food & Agriculture Org.

13. Groves CP. Domesticated and commensal mammals of Austronesia and their histories. In: The Austronesians. Ed: Bellwood P., Fox J.J. and Tryon D. The Australian National University Press, Canberra; 2006. pp.161–173.

14. Sudrajad, P. et al. (2020) 'An insight into the evolutionary history of Indonesian cattle assessed by whole genome data analysis', PLOS ONE, 15(11), p. e0241038. Available at: https://doi.org/10.1371/journal.pone.0241038.

15. Felius, M. et al. (2014) 'On the History of Cattle Genetic Resources', Diversity, 6(4), pp. 705–750. Available at: https://doi.org/10.3390/d6040705.

16. Konkruea, T. et al. (2017) ‘Genetic parameters and trends for daughters of imported and Thai Holstein sires for age at first calving and milk yield’, Agriculture and Natural Resources, 51(5), pp. 420–424. Available at: https://doi.org/10.1016/j.anres.2017.12.003.

17. Koonawootrittriron, S., Elzo, M.A. and Thongprapi, T. (2009) ‘Genetic trends in a Holstein×other breeds multibreed dairy population in Central Thailand’, Livestock Science, 122(2), pp. 186–192. Available at: https://doi.org/10.1016/j.livsci.2008.08.013.

18. Jattawa, D. et al. (2016) ‘Imputation Accuracy from Low to Moderate Density Single Nucleotide Polymorphism Chips in a Thai Multibreed Dairy Cattle Population’, Asian-Australasian Journal of Animal Sciences, 29(4), pp. 464–470. Available at: https://doi.org/10.5713/ajas.15.0291.

19. Ariyaraphong, N. et al. (2021) ‘High-Level Gene Flow Restricts Genetic Differentiation in Dairy Cattle Populations in Thailand: Insights from Large-Scale Mt D-Loop Sequencing’, Animals : an Open Access Journal from MDPI, 11(6), p. 1680. Available at: https://doi.org/10.3390/ani11061680.

20. Rhone, J.A., Koonawootrittriron, S. and Elzo, M.A. (2008) ‘Record keeping, genetic selection, educational experience and farm management effects on average milk yield per cow, milk fat percentage, bacterial score and bulk tank somatic cell count of dairy farms in the Central region of Thailand’, Tropical Animal Health and Production, 40(8), pp. 627–636. Available at: https://doi.org/10.1007/s11250-008-9141-6.

21. OECD (2022) OECD-FAO Agricultural Outlook 2022-2031. Paris: Organisation for Economic Co-operation and Development. Available at: https://www.oecd-ilibrary.org/agriculture-and-food/oecd-fao-agricultural-outlook-2022-2031_f1b0b29c-en

22. FAO, G. and I. (2020) Dairy’s Impact on Reducing Global Hunger: research summary: Global Agenda for Sustainable Livestock. Rome, Italy: FAO, GDP and IFCN. Available at: https://www.fao.org/publications/card/en/c/CB1198EN/

23. Foresight. The Future of Food and Farming (2011) Final Project Report. The Government Office for Science, London.

24. Bajželj, B. et al. (2014) ‘Importance of food-demand management for climate mitigation’, Nature Climate Change, 4(10), pp. 924–929. Available at: https://doi.org/10.1038/nclimate2353.

25. Jules Pretty, Zareen Pervez Bharucha, Sustainable intensification in agricultural systems, Annals of Botany, Volume 114, Issue 8, December 2014, Pages 1571–1596, https://doi.org/10.1093/aob/mcu205

26. J.E. Lombard et al (2019) "Proposed dairy calf birth certificate data and death loss categorization scheme", ScienceDirect, pp 4704-4712, https://doi.org/10.3168/jds.2018-15728.

27. Twomey, A.J., Berry, D.P., Evans, R.D. et al. Genome-wide association study of endo-parasite phenotypes using imputed whole-genome sequence data in dairy and beef cattle. Genet Sel Evol 51, 15 (2019). https://doi.org/10.1186/s12711-019-0457-7

28. A. Pacheco, et al (2021), 'Genetic parameters of animal traits associated with coccidian and nematode parasite load and growth in Scottish Blackface sheep', Animal, Volume 15, Issue 4, 2021, https://doi.org/10.1016/j.animal.2021.100185.

29. Thomson, S. et al. (2017) 'Bovine cryptosporidiosis: impact, host-parasite interaction and control strategies', Veterinary Research, 48(1), p. 42. Available at: https://doi.org/10.1186/s13567-017-0447-0.

30. Chappell, C.L. et al. (1996) 'Cryptosporidium parvum: intensity of infection and oocyst excretion patterns in healthy volunteers', The Journal of Infectious Diseases, 173(1), pp. 232–236. Available at: https://doi.org/10.1093/infdis/173.1.232.

31. Korich, D.G. et al. (1990) 'Effects of ozone, chlorine dioxide, chlorine, and monochloramine on Cryptosporidium parvum oocyst viability', Applied and Environmental Microbiology, 56(5), pp. 1423–1428. Available at: https://doi.org/10.1128/aem.56.5.1423-1428.1990.

32. Imre, K. and Dărăbuș, G. (2011) 'Distribution of Cryptosporidium species, genotypes and C. parvum subtypes in cattle in European countries', Sci Parasitol, 12, pp. 1–9.

33. Pinto, P. et al. (2021) 'Cross-Border Investigations on the Prevalence and Transmission Dynamics of Cryptosporidium Species in Dairy Cattle Farms in Western Mainland Europe', Microorganisms, 9(11), p. 2394. Available at: https://doi.org/10.3390/microorganisms9112394.

34. Diaz, E. et al. (2003) 'Epidemiology and control of intestinal parasites with nitazoxanide in children in Mexico', The American Journal of Tropical Medicine and Hygiene, 68(4), pp. 384–385.

35. 42. Barton, N.H. (2010) 'Genetic linkage and natural selection', Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1552), pp. 2559–2569. Available at: https://doi.org/10.1098/rstb.2010.0106.

36. Gillespie, J.H. (2001) 'Is the Population Size of a Species Relevant to Its Evolution?', Evolution, 55(11), pp. 2161–2169. Available at: https://doi.org/10.1111/j.0014-3820.2001.tb00732.x.

37. Molecular Cloning: A Laboratory Manual, 3rd ed., Vols 1,2 and 3 J.F. Sambrook and D.W. Russell, ed., Cold Spring Harbor Laboratory Press, 2001, 2100 pp., soft cover | Sigma-Aldrich (no date). Available at: http://www.sigmaaldrich.com/

38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81. 46. http://pngu.mgh.harvard.edu/purcell/plink/

39. Mi, H. et al. (2019) 'Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)', Nature Protocols, 14(3), pp. 703–721. Available at: https://doi.org/10.1038/s41596-019-0128-8.

40. Decker, Jared Egan et al. (2015), Data from: Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle, Dryad, Dataset, https://doi.org/10.5061/dryad.th092

41. Coordinate remapping service: NCBI Available at: www.ncbi.nlm.nih.gov/genome/tools/remap

42. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) 'Inference of population structure using multilocus genotype data', Genetics, 155(2), pp. 945–959. Available at: https://doi.org/10.1093/genetics/155.2.945.

43. Earl, Dent A. and vonHoldt, Bridgett M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources vol. 4 (2) pp. 359-361 doi: 10.1007/s12686-011-9548-7

44. Gilbert, K.J. (2016) 'Identifying the number of population clusters with structure: problems and solutions', Molecular Ecology Resources, 16(3), pp. 601–603. Available at: https://doi.org/10.1111/1755-0998.12521.

45. Pickrell, J.K. and Pritchard, J.K. (2012) 'Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data', PLOS Genetics, 8(11), p. e1002967. Available at: https://doi.org/10.1371/journal.pgen.1002967.

46. Fitak, R.R. (2021) 'OptM: estimating the optimal number of migration edges on population trees using Treemix', Biology Methods and Protocols, 6(1), p. bpab017. Available at: https://doi.org/10.1093/biomethods/bpab017.

47. Bourgeois, Stéphanie, Jenny Kaden, Helen Senn, Nils Bunnefeld, Kathryn J. Jeffery, Etienne F. Akomo-Okoue, Rob Ogden, and Ross McEwing. 'Improving Cost-Efficiency of Faecal Genotyping: New Tools for Elephant Species'. PLOS ONE 14, no. 1 (30 January 2019): e0210811. https://doi.org/10.1371/journal.pone.0210811.

48. Schultz, Anthony J., Romane H. Cristescu, Bethan L. Littleford-Colquhoun, Damian Jaccoud, and Céline H. Frère. 'Fresh Is Best: Accurate SNP Genotyping from Koala Scats'. Ecology and Evolution 8, no. 6 (2018): 3139–51. https://doi.org/10.1002/ece3.3765

49. Snyder-Mackler, N. et al. (2016) 'Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples', Genetics, 203(2), pp. 699–714. Available at: https://doi.org/10.1534/genetics.116.187492.

50. Maudet, C., G. Luikart, D. Dubray, A. Von Hardenberg, and P. Taberlet. 'Low Genotyping Error Rates in Wild Ungulate Faeces Sampled in Winter'. Molecular Ecology Notes 4, no. 4 (2004): 772–75. https://doi.org/10.1111/j.1471-8286.2004.00787.x.

51. Lahiri, Debomoy K., Steve Bye, John I. Nurnberger, Marion E. Hodes, and Margaret Crisp. 'A Non-Organic and Non-Enzymatic Extraction Method Gives Higher Yields of Genomic DNA from Whole-Blood Samples than Do Nine Other Methods Tested'. Journal of Biochemical and Biophysical Methods 25, no. 4 (1 January 1992): 193–205. https://doi.org/10.1016/0165-022X(92)90014-2.

52. Neary, M. T., J. M. Neary, G. K. Lund, F. B. Garry, T. N. Holt, T. J. Mohun, and R. A. Breckenridge. 'Technical Note: A Comparison of DNA Collection Methods in Cattle and Yaks1'. Journal of Animal Science 92, no. 9 (1 September 2014): 3811–15. https://doi.org/10.2527/jas.2013-7445.

53. Nayee, N. et al. (2018) 'Suitability of existing commercial single nucleotide polymorphism chips for genomic studies in Bos indicus cattle breeds and their Bos taurus crosses', Journal of Animal Breeding', 135(6), pp. 432–441. Available at: https://doi.org/10.1111/jbg.12356

54. Elliott, D. A., & Clark, D. P. (2000). Cryptosporidium parvum induces host cell actin accumulation at the host-parasite interface. Infection and immunity, 68(4), 2315–2322. https://doi.org/10.1128/IAI.68.4.2315-2322.2000

55. Vélez, J. et al. (2021) 'Metabolic Signatures of Cryptosporidium parvum-Infected HCT-8 Cells and Impact of Selected Metabolic Inhibitors on C. parvum Infection under Physioxia and Hyperoxia', Biology, 10(1), p. 60. Available at: https://doi.org/10.3390/biology10010060.

56. Chen, X.-M. et al. (2004) 'Cdc42 and the actin-related protein/neural Wiskott-Aldrich syndrome protein network mediate cellular invasion by Cryptosporidium parvum', Infection and Immunity, 72(5), pp. 3011–3021. Available at: https://doi.org/10.1128/IAI.72.5.3011-3021.2004.

57. Matsubayashi, M. et al. (2010) 'Morphological changes and viability of Cryptosporidium parvum sporozoites after excystation in cell-free culture media', Parasitology, 137(13), pp. 1861–1866. Available at: https://doi.org/10.1017/S0031182010000685.

58. Li, T. et al. (2021) 'Comparative proteomics reveals Cryptosporidium parvum manipulation of the host cell molecular expression and immune response', PLOS Neglected Tropical Diseases, 15(11), p. e0009949. Available at: https://doi.org/10.1371/journal.pntd.0009949.

59. Kaluz, S. et al. (2009) 'Transcriptional control of the tumor- and hypoxia-marker carbonic anhydrase 9: a one transcription factor (HIF-1) show?', Biochimica et biophysica acta, 1795(2), pp. 162–172. Available at: https://doi.org/10.1016/j.bbcan.2009.01.001.

60. Borenshtein, D. et al. (2009) 'Decreased Expression of Colonic Slc26a3 and Carbonic Anhydrase IV as a Cause of Fatal Infectious Diarrhea in Mice', Infection and Immunity, 77(9), pp. 3639–3650. Available at: https://doi.org/10.1128/IAI.00225-09

61. O'Hara, S. and Lin, J. (2006) 'Accumulation of tropomyosin isoform 5 at the infection sites of host cells during Cryptosporidium invasion', Parasitology research, 99, pp. 45–54. Available at: https://doi.org/10.1007/s00436-005-0117-4.

62. Owen, L.M. et al. (2022) 'The C-terminal actin-binding domain of talin forms an asymmetric catch bond with F-actin', Proceedings of the National Academy of Sciences of the United States of America, 119(10), p. e2109329119. Available at: https://doi.org/10.1073/pnas.2109329119.

63. Finlay, B.B. et al. (1992) 'Cytoskeletal composition of attaching and effacing lesions associated with enteropathogenic Escherichia coli adherence to HeLa cells.', Infection and Immunity, 60(6), pp. 2541–2543.

64. Finlay, B.B., Ruschkowski, S. and Dedhar, S. (1991) 'Cytoskeletal rearrangements accompanying salmonella entry into epithelial cells', Journal of Cell Science, 99 (Pt 2), pp. 283–296. Available at: https://doi.org/10.1242/jcs.99.2.283.

65. Watarai, M. et al. (1997) 'rho, a small GTP-binding protein, is essential for Shigella invasion of epithelial cells', The Journal of Experimental Medicine, 185(2), pp. 281–292. Available at: https://doi.org/10.1084/jem.185.2.281.

66. Cekan, S.Z. (2004) 'Methods to find out the expression of activated genes', Reproductive biology and endocrinology : RB&E, 2, p. 68. Available at: https://doi.org/10.1186/1477-7827-2-68.

67. Mei, C. et al. (2021) 'Insights into adaption and growth evolution: a comparative genomics study on two distinct cattle breeds from Northern and Southern China', Molecular Therapy. Nucleic Acids, 23, pp. 959–967. Available at: https://doi.org/10.1016/j.omtn.2020.12.028.

68. Cai, Y. et al. (2018) 'Maternal genetic and phylogenetic characteristics of domesticated cattle in northwestern China', PLOS ONE, 13(12), p. e0209645. Available at: https://doi.org/10.1371/journal.pone.0209645.

69. Wangkumhang, P. et al. (2015) 'Genetic analysis of Thai cattle reveals a Southeast Asian indicine ancestry', PeerJ, 3, p. e1318. Available at: https://doi.org/10.7717/peerj.1318.

70. H Martojo (2012). Indigenous Bali Cattle is Most Suitable for Sustainable Small Farming in Indonesia. , 47(Supplement s1), 10–14. doi:10.1111/j.1439-0531.2011.01958.x

71. Decker, J.E. et al. (2014) 'Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle', PLOS Genetics, 10(3), p. e1004254. Available at: https://doi.org/10.1371/journal.pgen.1004254.

72. Lawson, D.J., van Dorp, L. and Falush, D. (2018) 'A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots', Nature Communications, 9(1), p. 3258. Available at: https://doi.org/10.1038/s41467-018-05257-7.

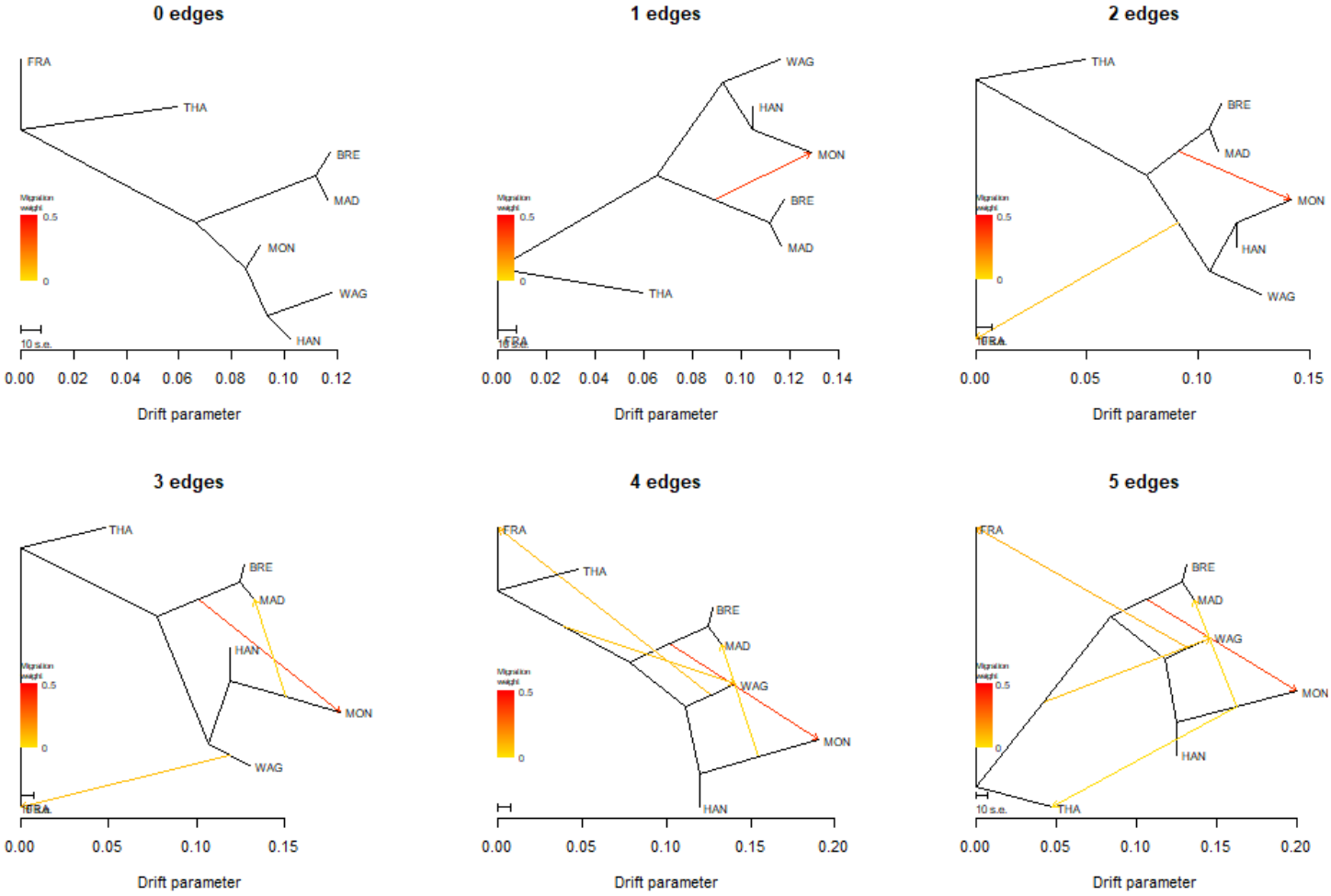
Appendix 1 (additional data).

Appendix table 1.

Country	Farm Name	Age	Bristol Score	C.Parvum presence.
Cyprus	Strouthos	Unknown	5	Negative
	41010		5	
	41008		5	
	21406		5	
	21407		4	
	41010		4	Positive
	41008		7	
	21407		5	
	21406		6	
	Strouthos		7	
Netherlands	ADR/RIC	1 day	4	Negative
	KEI/JOH	3 days	6	
	HER/FRI	7 days	6	
	KRO/LAR	17 days	7	
	WEZ/MAR	8 days	7	
	ADR/RIC	6 days	7	Positive
	HER/FRI	10 days	7	
	POP/PAU	23 days	7	
	WEZ/MAR	16 days	5	
	KRO/LAR	8 days	7	
France	2FRAUXSEP	23 days	Unknown	Negative
	2FRFLAING	4 years		
	2FRCVEDUR	Unknown		
	2FRDELGAF			
	2FRDEUCAT	35 days		

RoI coordinate including the +/-500bp positions (Chr : BP position)	
1:10033802:10133802	1:10033802:10133802
2:14734977:14834977	2:14734977:14834977
2:44427952:44527952	2:44427952:44527952
3:29258354:29358354	3:29258354:29358354
6:58862805:58962805	6:58862805:58962805
8:59852102:59952102	8:59852102:59952102
8:112420954:112520954	8:112420954:112520954
11:78429299:78529299	

Appendix table 2.



Appendix Figure 1.

Appendix 2 (coding and scripts).

Appendix 2.1 PLINK (RStudio) Coding.

Initial Steps

- Clear workspace
- Set working directory

```
#Load (or download) relevant packages#
library ("tidyverse")
library ("qqman")
library("ggplot2")
library("Polychrome")
library("biomaRt")
```

Preparing files, quality control and file conversion

NOTE: –cow and –chr-set 29 can be used interchangeably.

```
#  Quality Control criteria
## Missingness per SNP; --geno
## Missingness per individual; --mind
## Minor Allele Frequency; --maf
## Hardy-Weinberg threshold;--hwe

#Analysis
## Estimates of expected heterozygosity --het
## Allele/genotype frequency report --freq
## Hardy-Weinberg equilibrium exact test p-value report --hardy

# QC2 - repeat separately for Crypto cattle and then Asian cattle. (exlu.
Additional Dataset, which is instead filterered according to QC1).

system ("plink --bfile FILE --cow --autosome --geno 0.2 --mind 0.4 --maf
0.05 --hwe 0.0000001 --allow-no-sex --nonfounders --hardy --freq --het --m
ake-bed --out Filtered_FILE")

# Information for missingness per individual:
system ("plink --bfile Filtered_FILE --cow --autosome --missing --out Miss
ingness_FILE")
```

Creating Subsets, wherein ‘FAM’ is:

BRE, MAD, WAG, HAN, MON, THA for the Asian samples. NEG and POS for the Crypto samples. (‘FAM.txt’ is a .txt file containing only the FAM identifier, eg BRE.txt).
The end of this document provides further details on generating subsets.

```
system("plink --bfile Filtered_FILE --cow --keep-fam FAM.txt --make-bed --
out FAM_FILE")
# Merge to generate sets - a set for the Asian samples and a set for the C
rypto samples.
system("plink --bfile FAM_FILE --cow --autosome --bmerge OTHER_FAM_FILE --
nonfounders --allow-no-sex --make-bed --out SET_FILE")

# Merged sets may contain duplicate positions/SNPs. To remove:
system(paste0("plink --bfile SET_FILE --cow --autosome ",
              "--list-duplicate-vars ids-only suppress-first --nonfounders
 ",
              "--allow-no-sex --make-bed --out SET_FILE_duplicates"))


system(paste0("plink --bfile SET_FILE --cow --autosome ",
              "--exclude SET_FILE_duplicates.dupvar --nonfounders ",
              "--allow-no-sex --make-bed --out SET_FILE_dupRemoved"))
```

ASIAN ANALYSIS

```
# Genetic distances between individuals
system("plink --cow --allow-no-sex --nonfounders --bfile ASIAN_FILE_dupRem
oved --distance-matrix --out DistAsianFile")

# Load data frame
dist_populations<-read.table("DistAsianFile.mdist",header=F)

# Extract breed names
fam <- data.frame(famids=read.table("AsianCryptoFile.mdist.id")[,1])

# Extract indvidual names
famInd <- data.frame(IID=read.table("AsianCryptoFile.mdist.id")[,2])

# Perform PCA using the cmdscale function - and produce summary
mds_populations <- cmdscale(dist_populations,eig=T,5)
summary(mds_populations)

# Extract the eigen vectors and values
eigenvec_populations <- cbind(fam,famInd,mds_populations$points)

# Proportion of variation captured by each eigen vector - and produce summ
ary
eigen_percent <- round(((mds_populations$eig)/sum(mds_populations$eig))*10
0,2)
view(eigen_percent)
```

```
# Produce PCA plot
ggplot(data = eigenvec_populations) +
  geom_point(mapping = aes(x = `1`, y = `2`,color = famids), show.legend =
 TRUE ) +
  geom_hline(yintercept = 0, linetype="dotted") +
  scale_color_brewer(palette="Set1") +
  geom_vline(xintercept = 0, linetype="dotted") +
  labs(title = "PCA Asian Breeds",
       x = paste0("Principal component 1 (",eigen_percent[1]," %)"),
       y = paste0("Principal component 2 (",eigen_percent[2]," %)")) +
  theme_minimal()

#Compute FST
system("plink --cow --bfile ASIAN_FILE_dupremoved --fst --family --out ASI
AN_FILE_FstResults")

#Generating an input file suitable for STRUCTURE
system("plink --bfile ASIAN_FILE_dupremoved --chr-set 29 --recode structur
e --out ASIAN_Structure")
```

CRYPTOSPORIDIUM ANALYSIS

```
# Compute FST
system("plink --cow --bfile CRYPTO_FILE_dupremoved --fst --family --out CR
YPTO_FILE_FstResults")

# Visualise FST Results
# For this, the suggestive line was set to 0.45; in highly differentiated
populations (i.e different breeds), this should be raised to 0.5.

data <- read.delim("CRYPTO_FILE_fstResults.fst") %>%
  drop_na()
manhattan(data,  col = palette36.colors(n=29), chr = "CHR", bp = "POS", p
= "FST", suggestiveline = 0.45, logp = F)   #setting suggestiveline for vi
sualisation

# Extracting gene content from genomic regions
segmentsBiomart <- data %>%
  filter(FST >= 0.45) %>%
  mutate(RoI = paste(CHR,POS-50000,POS+50000,sep = ":")) %>% ##identifying
 genes +/-50,000 bp RoI
  dplyr::select(RoI) %>%
  t()

write_delim(as.data.frame(t(segmentsBiomart)), "RoI4BioMart50000.txt", col
_names = F,delim = ",")

browseVignettes("biomaRt")

listMarts(host="https://www.ensembl.org")
ensembl <- useMart("ensembl")
datasets <- listDatasets(ensembl) ## to show names of all databases
```

```
ensembl = useMart(biomart="ENSEMBL_MART_ENSEMBL",dataset="btaurus_gene_ens
embl") ##using btaurus gene ensembl

attributes = listAttributes(ensembl)

filters = listFilters(ensembl)

biomaRt_results=getBM(attributes = c("ensembl_gene_id", "external_gene_nam
e", "gene_biotype",
                                     "go_id", "name_1006", "definition_100
6", "go_linkage_type", "namespace_1003",
                                     "chromosome_name", "start_position",
"end_position","description"),
                      filters = c("chromosomal_region","biotype"),
                      values =list(chromosomal_region=segmentsBiomart,biot
ype="protein_coding"),
                      mart =ensembl)
write_delim(biomaRt_results, "resultsFromBioMart50000.txt", delim = "\t")
```

= The resulting .txt document contains the gene identifiers for input into PANTHER (or other gene-ontology classification softwares)

Misc. coding - creating subsets

```
cattle <- read.table("Filtered_FILE.fam")

# select Cattle from target countries and apply country label
countryLabels <- cattle %>%
  mutate(country = case_when(
    substr(V1,1,3) == "FAM1" ~ "Subset1",
    substr(V1,1,3) == "FAM2" ~ "Subset1",
    substr(V1,1,3) == "FAM3" ~ "Subset2",
    substr(V1,1,3) == "FAM4" ~ "Subset2",
  )) %>%
  drop_na() %>%
  select(V1, V2, country) %>%
  write_delim("FAM_SETs.txt")

# Now the FAM is defined instead by a given SET- keep only a wanted SET in
 new PLINK files #

system(str_c("plink --bfile Fi --chr-set 29 --autosome ",
             "--keep KEEP_SETS.txt --nonfounders ",
             "--within KEEP_SETS.txt ",
             "--out NEW_SETs"))
```

Appendix 2.2 TreeMix input scripts.

Initial Steps

- Clear workspace
- Set working directory

The end of this document provides further information, detail and coding on merging SNP datasets and filtering the data.

Linkage Pruning

TreeMix does not like missing data. Remove sites with missing data by performing linkage-pruning:

```
system ("plink --cow --file Filtered_FILE --indep-pairwise 50 10 0.1 --now
eb --out Filtered_FILE --silent")

#Keeping only those SNPs not in LD:
system ("plink --file Filtered_FILE --extract Filtered_FILE.prune.in --cow
 --recode --out Pruned_FILE")

# Generating plink files

system ("plink --file Pruned_FILE --cow --make-bed --out Pruned_FILE")
system ("plink --bfile Pruned_FILE --recode --out Pruned_FILE --allow-no-s
ex --cow")
```

Generate a .clust file:

```
system ("plink --file Pruned_FILE --autosome --cow --cluster --out Cluster
_FILE")
```

This will output 3 files; a .cluster1, .cluster2, .cluster3 file. The .cluster2 file is required as a .CLUST file. Rename the .cluster2 file as; Cluster_FILE.CLUST to produce a .CLUST file.

Generate a .frq.strat file:

```
system ("plink --file Pruned_FILE --freq --missing --within Cluster_FILE.c
lust --out Input_FILE --allow-no-sex --cow")
```

Generating TreeMix Files

- Download TreeMix (and Python3)

Download and run 'ConvertPlink2TreeMix' script (available here: https://github.com/thomnelson/tools/blob/master/plink2treemix.py)

This script uses: - Input_FILE.frq.strat - plink2treemix.py(https://github.com/thomnelson/tools/blob/master/plink2treemix.py)

Save the output as a .treemix.frq.gz file (Input_FILE.treemix.frq.gz)

The file should read as: [Population headings] [Data]

Running TreeMix

This Input_FILE.treemix.frq.gz can be input into TreeMix via command line:

```
 for i in {0..5}
 do
 treemix -i Input_FILE.treemix.frq.gz -m $i -o FILE.$i -root ROOT -bootstr
ap -k 500 -noss > treemix_${i}_log & done

# -root choose the position of the root/outgroup
# -m the number of migration events to investigate
# -bootstrap generate a bootstrap replicate, with -k 500 determining resam
pling blocks of 500 SNPs (For judging the confidence in a given tree topol
ogy, it is often of interest to generate a bootstrap replicate)
# -noss turns off sample size correction (useful for smaller sample sizes)
```

Save all output files to appropriate folder.

##Visualising TreeMix output files. Download plotting_funcs.R and place in appropriate folder, from https://github.com/joepickrell/pophistory-tutorial/blob/master/example2/plotting_funcs.R

```
setwd("/appropriatefolder") # this needs to be adjusted
prefix="FILE" #This is the prefix denoted by -o in TreeMix

library(RColorBrewer)
library(R.utils)
source("plotting_funcs.R") #adding the path

par(mfrow=c(1,1)) #Can be altered (e.g 2,3)
for(edge in 0:5){ #Migration edges to plot
  plot_tree(cex=0.8,paste0(prefix,".",edge))
  title(paste(edge,"edges"))
}
```

Using OptM to estimate optimal Migration Edges

From: https://cran.r-project.org/web/packages/OptM/readme/README.html

```
library(SiZer)
library(OptM)

folder <- system.file("extdata", package = "OptM")
test.optM = optM(folder)
plot_optM(test.optM, method = "Evanno")
```

#Misc. coding - Merging datasets and filtering. Merging datasets: The Asian-breed dataset has a smaller range of SNPs than the European-breed dataset (due to additional Asian breeds based on alternate genome assembly and 50k SNP CHP.

```
## Merging datasets based on common SNPs for analysis.
map2 = read.delim("Dataset1.map", header=F, quote="")
map1 = read.delim("Dataset2.map", header=F, quote="")
common.snps = which(map2$V2 %in% map1$V2)
write.table(map2$V2[common.snps], file="list.snps", sep="\t", col.names=F,
```

```
 row.names=F, quote=F )

system( "plink --bfile Dataset1 --cow --extract list.snps --make-bed --out
 Dataset1List")
system ("plink --bfile Dataset2 --cow --extract list.snps --make-bed --out
 Dataset2List")
system ("plink --bfile Dataset1List --bmerge Dataset2.bed Dataset2List.bim
 Dataset2List.fam --make-bed --cow --out MergedDataset")


## Merged sets may contain duplicate positions/SNPs. To remove:
system(paste0("plink --bfile MergedDataset --cow --autosome ",
              "--list-duplicate-vars ids-only suppress-first --nonfounders
 ",
              "--allow-no-sex --make-bed --out MergedDataset_duplicates"))


system(paste0("plink --bfile MergedDataset --cow --autosome ",
              "--exclude MergedDataset_duplicates.dupvar --nonfounders ",
              "--allow-no-sex --make-bed --out MergedDataset_dupRemoved"))

# Filtering the data
system ("plink --bfile MergedDataset_dupRemoved --cow --autosome --geno 0.
4 --mind 0.4 --maf 0.00005 --hwe 0.0000001 --allow-no-sex --nonfounders --
hardy --freq --het --make-bed --out Filtered_FILE")
```