THE UNIVERSITY *of* EDINBURGH

The Epidemiology of East Coast Fever in Smallholder Livestock Systems of East Africa

THE UNIVERSITY *of* EDINBURGH

Sofia Findlay-Pacheco

Master of Science by Research
2025

Declaration

I declare that this dissertation was written by myself, and the work contained within it is my own unless explicitly stated. This work has not been submitted for any other degree/qualification.

Sofia Findlay-Pacheco

Lay summary

The health and production rates of livestock play an important role in supporting livelihoods in Africa. Livestock is relied upon for nutrition, employment and foreign trades. However, infectious diseases, particularly those caused by tiny blood-borne pathogens known as haemopathogens, can lower productivity by as much as 25%. Haemopathogens, including species like *Theileria parva, Anaplasma marginale, Ehrlichia,* and *Babesia* species, cause a wide array of clinical signs such as fever, anemia, organ dysfunction, and death.

East Coast fever (ECF) is a devastating infectious disease caused by the *T. parva* pathogen and spread by ticks affecting Sub-Saharan Africa. Every year ECF kills more than 1.1 million cattle and causes economic losses exceeding US $300 million. Cattle that get ECF often show signs like swollen lymph nodes, fever, and extreme tiredness and many do not survive. Farmers' livelihoods are severely impacted by ECF infections through reduced milk production, fertility issues and increased veterinary costs.

Some specific cattle breeds have shown natural resistance to ECF. This study examines this further through the impact of the *FAF1B* gene. This gene has been found to be associated with inherited tolerance to cases of ECF originating from buffalo populations. Understanding its influence could allow for future breeding programs to produce resistant/tolerant cattle. Control of ECF remains challenging; acaricides are costly, environmentally harmful, and increasingly ineffective due to acaricide-resistant ticks. While the current ECF vaccine referred to as the Muguga cocktail ITM vaccine is effective, it’s less reliable in places with buffalo, which carry a different strain of the pathogen.

A further challenge when studying disease in livestock is the possibility of co-infections, where animals are infected with multiple pathogens at the same time. These interactions can alter immune responses, reduce vaccine efficacy, and obscure diagnosis due to overlapping clinical signs. In poorer areas, diagnosing these infections is even harder due to a lack of access to advanced tools like molecular testing.

In order to carry out this analysis data from the 2007–2009 IDEAL study which tracked nearly 550 calves in Kenya over their first year of life was analysed. The study collected information about the calves' health, farm conditions, infections, and causes of death, along with lab test results and genetic data. Using R, a program used to analyze large amounts of data and run

statistical tests, the data was explored specifically looking at how genetics, co-infections, and infection timing influenced which calves lived or died from ECF.

This study showed the promising presence of tolerant genetic profiles against ECF, calves who were homozygous for the T allele of the specified gene were more likely to survive the disease. The study also found that calves exposed to milder, less dangerous species of *Theileria* prior to getting infected with the deadly strain were more likely to survive. Understanding these relationships could guide improved vaccination strategies and inform broader vector borne infectious disease control efforts in order to reduce both mortality and economic losses.

Abstract

The health and productivity of livestock are pivotal in rural livelihoods across sub-Saharan Africa, providing nutrition, income and employment. However, infectious diseases, particularly those caused by haemopathogens, pose a significant threat to their health and productivity. Haemopathogens including *Theileria parva*, *Anaplasma marginale*, *Ehrlichia ruminatium*, and *Babesia* spp. infect the host's bloodstream, causing signs like fever, anaemia, organ dysfunction and death. East Coast fever (ECF), caused by *T. parva* and transmitted by the tick *Rhipicephalus appendiculatus*, is among the deadliest of these infections and causes the development of lymphoproliferative syndrome and the significant reduction in milk production, fertility and growth. Co-infections are common in endemic regions with cattle often harbouring multiple pathogens concurrently which may result in either synergistic or antagonistic effects on disease progression. Co-infections can result in altered immune responses, misdiagnosis, and reduced treatment efficacy. Certain co-infections, specifically other *Theileria* species may offer protective effects when a calf contracts ECF. Recent research has identified potential heritable genetic tolerance to ECF, particularly involving a variant in the *FAF1B* gene which regulates immune-mediated apoptosis.

This study aims to investigate the epidemiology, transmission, co-infection dynamics, and inherited tolerance of haemopathogen diseases in East African smallholder cattle systems, with a specific focus on ECF. Data from the 2007–2009 Infectious Diseases of East African Livestock (IDEAL) study, including novel Illumina MiSeq pathogen profiles, from a haemobiome tool developed by the University of Edinburgh, Centre for Tropical Livestock Genetics and Health (CTLGH) and cattle genetic data, were used to conduct statistical analyses. Results were used to assess the influencers and predictors of calf survival and disease outcome.

Calves homozygous for the *FAF1B (FAS-associated factor 1)* T allele (TT) exhibited a reduced risk in mortality attributed to ECF, supporting a potential role of genetically mediated tolerance. Additionally, prior exposure to less pathogenic *Theileria* species *(Theileria-mutans and/or Therileria-velifera)* before infection with *T. parva* was statistically associated with a higher chance of survival, implying that this co-infection plays a protective role. The findings aid in informing effective disease management strategies, including selective breeding for tolerant genotypes and harnessing the potential of co-infections to mitigate disease severity ultimately contributing to reduced cattle mortality and greater economic resilience for smallholder farmers.

Table of contents

Glossary

ECF : East Coast fever

IDEAL : Infectious Diseases of East African Livestock

CTLGH : Centre for Tropical Livestock Genetics and Health

FAF1 : FAS-associated factor 1

qPCR : Quantitative Polymerase Chain Reaction

CTL : Cytotoxic T-lymphocyte cell

ELISA : Enzyme-linked immunosorbent assay

DNA : Deoxyribonucleic acid

RLB : Reverse Line Blot hybridization

AEZ : Agro-ecological zones

IFNs : Immunity driven by cytokines like interferons

SNP : Single nucleotide polymorphism

MID : Multiplex identifier tags

1. Literature review

1.1 Haemopathogens in smallholder livestock systems

The health and productivity of livestock play a pivotal role in supporting livelihoods in Low and middle income countries, in particular in Sub-Saharan Africa where smallholder systems dominate the agricultural scene. Livestock therefore, while providing the main source of income for most families, are equally relied upon for nutrition in the form of protein for the human diet, employment and domestic exchanges (Nene et al., 2016; Upton, 2004). Currently, Africa holds one third of global livestock populations, with Kenya accounting for 21.9 million cattle as of 2023. This sector is seeing increased investments due to growing urbanization and income growth (Balehegn et al., 2021; *Cattle Population by Country*, 2025).

Diseases that pose a threat to the wellbeing and productivity of livestock are therefore a considerable challenge, of which haemopathogens make up a large proportion. Haemopathogens are pathogens that colonise the bloodstream of the host, in particular this study will concentrate on *Theileria, Anaplasma*, *Ehrlichia and Babesia (Stuen, 2020)*. Haemopathogens are particularly impactful due to the range of pathologies they induce in livestock, such as anaemia, immunosuppression, and organ dysfunction. These underlying disease processes manifest clinically as reductions in appetite, fever, organ dysfunction or in severe cases death (Maharana et al., 2016). These direct effects on the animals' health will impact a farmer's income through a reduction in productivity via reduced milk production, fertility and animal growth rates, and ultimately the death of their livestock (Stuen, 2020). Globally, livestock disease is responsible for a 25% decrease in productivity (Džermeikaitė et al., 2024). In addition, farmers face financial pressures from increased expenditure on medication, veterinary services, and carcass disposal (Džermeikaitė et al., 2024).

Beyond farm-level impacts, animal disease also has significant environmental consequences. A reduction in cattle productivity will often require larger herds to maintain the same level of milk or meat output, leading to an increase in greenhouse gas emissions (*FAO,* 2022, *How Improved Livestock Health Can Reduce GHG Emissions*, 2024; Maharana et al., 2016; Minjauw & Mcleod, 2003). It’s estimated that an unhealthy cow can increase greenhouse gas emissions of milk by up to 25% per ton of milk. The use of chemical treatments and their impact on the land will have their own carbon footprint and impact other untargeted areas of our ecosystem such as other plants, animals and invertebrates. For example, acaricides are typically administered directly to cattle rather than applied to the land, but residues can enter the environment through animal waste, wash off or other pathways. (*FAO*, 2022, *How Improved Livestock Health Can Reduce GHG Emissions*, 2024; Soliman et al., 2023)

Haemopathogens are most commonly transmitted via vectors including ticks, tsetse flies and mosquitos, oftentimes serving as both intermediates and reservoirs. The favourable warm and humid ecological regions found in East Africa allow ticks to thrive, however transmission to otherwise previously uninfected areas is dependent on factors such as temperature, humidity and vegetation (Soliman et al., 2023). Population surges in ticks are commonly seen following large rainfalls as it supports vegetation and therefore the habitats ticks live in. While dry and hot seasons have been shown to adversely impact tick questing activity as the exposure to these conditions will increase their risk of desiccation (Dantas-Torres, 2015; Leal et al., 2020).

These surges can be seen by the geographical distribution and abundance of species. The expansion of tick habitats into previously unsuitable areas will be seen in regions where warming temperatures and milder winters are the result of climate change. This has been seen for example in the distribution of the castor bean tick *Ixodes ricinus* where a 30 year study in Sweden displayed a clear expansion from the tick into northerly regions of the country, and the ticks coverage of the northern areas (north of 60°N) doubled from 12.5% to 26.8% between the 1990s and 2008 (Dantas-Torres, 2015). However these expansions are not unidirectional and regions of excessive heat as a result of climate change, may become less favourable habitats and result in a decline in tick population as the area becomes too dry to support survival and life cycles. Predictions have been made in South Africa for reduced habitats for tick species such as *Rhipicephalus decoloratus, Amblyomma hebraeum, Rhipicephalus appendiculatus* and *Hyalomma truncatum* as a result of a 2°C temperature rise (Ebert & Becker, 2025). Human activities such as deforestation and overgrazing further exacerbate these dynamics by altering habitats and increasing contact between vectors, livestock, and wildlife, thereby influencing patterns of disease transmission (Abdullah et al., 2019).

This report will concentrate on Smallholder Livestock systems that account for up to 80 percent of the food production in areas of Sub-Saharan Africa and Asia (*IFAD*, 2013.). The characteristics of a smallholder farm are small-scale farms that are between 1 to 10 hectares, they are most commonly family-focused and primarily used to meet a household's needs, however, they may also sell their products for extra income (*Smallholders and Family Farmers*, 2013). As part of these systems the farmers will tend to rely upon natural grazing, crop residues, or foraged fodder to sustain their animals. This reliance on locally available feed resources reflects the resource constraints that also explain the minimal use of medicines, or advanced farming technologies (Pfeiffer et al., 2022).

Such characteristics shape how haemopathogens impact these farms. On the one hand the dispersed and small group nature of smallholder farms will reduce the transmission potential in haemopathogens that spread primarily through direct contact, for example this will be the case in disease such as contagious bovine pleuromonia where close and frequent animal-to-animal contact is the main driver of outbreaks (Fenta et al., 2024). While mitigating for some diseases, in the case of vector-borne haemopathogens the smallholder characteristics will heighten the vulnerability of cattle. Cattle grazing in communal pastures and whose wider grazing areas will lead them to come into greater contact with tick habitats and neighboring herds will lead to a

greater risk of exposure to infected vectors for the cattle (Ekwem et al., 2021). This on top of limited access to veterinary services and preventive measures further constrains farmers' ability to manage outbreaks, meaning that the consequences of vector-borne infections can be particularly severe in these systems (Arvidsson et al., 2022).

Smallholder farming systems are diverse, encompassing dairy, meat and mixed enterprises. The distribution of these systems across Africa is closely linked to local disease risks. For example dairy farming is less common in areas with high endemicity ECF, as research has shown that exotic or crossbread cattle with higher proportion of European genetics (those associated with higher milk yield) have shown higher susceptibility to *T. parva* infections (Murray et al., 2013). A further limitation in these systems is that smallholder farms are generally situated in resource-poor settings where disease surveillance is often limited, leading to weak systematic control efforts. As a result, epidemiological data are scarce, making it difficult to monitor disease burdens or detect emerging trends (Garcia et al., 2022; Callaby et al., 2020).

1.2 Co-infections

A further challenge when studying disease in livestock is the possibility of co-infections, these occur when one livestock host is simultaneously infected by multiple pathogens including bacteria, viruses, protozoa, and helminths (Thumbi et al., 2014). The interaction between these species can influence a host's immune response and alter the trajectory of the disease. These interactions can be categorized into synergistic or antagonistic effects, a synergistic impact will occur when two haemopathogens interact and cause the amplification of the severity of the clinical signs seen and therefore the progression of the disease (Akoolo et al., 2022).

This is especially relevant in cases of vaccination aimed at protecting cattle from a specific infection as it impacts the hosts ability to form long-lasting memory responses, however it is not restricted to vaccines and can also be seen affecting natural infection outcomes (Akoolo et al., 2022). A mechanism behind synergistic impacts is immune skewing, a mechanism by which a co-infection from Th1 and Th2 response inducing pathogens can greatly impact disease progression. In the context of this paper this can be seen in cattle suffering from ECF however also co-infected with Helminths (McNeilly & Nisbet, 2014). ECF suppression is reliant on a Th1 response from the host, this means cell-mediated immunity driven by cytokines like interferons (IFNs)-γ and IL-2 and which activates macrophages and cytotoxic T-cells. The body's ability to mount an immune response to helminths on the other hand relies on a Th2 response, this is reliant on humoral/antibody-mediated immunity driven by cytokines like IL-4, IL-5, IL-13 and the promotion of B-cell activation and antibody production. These two arms of the immune system are partly antagonistic and so a dominant Th2 response will downregulate a Th1 response (Rincón & Flavell, 1997). This can mean vaccines dependent on strong Th1 responses may be less effective (Akoolo et al., 2022; McNeilly & Nisbet, 2014).

Antagonistic interactions are when 2 pathogens infect a host and reduce the severity of infection or replication rate of the other (Thumbi et al., 2014). These interactions are important to identify and understand as antagonistic relations where a previously lethal haemopathogen is being

suppressed and therefore less impactful, this can be a benefit to the smallholder farms. A classic historical example of an antagonistic relationship is seen between cowpox and smallpox viruses. Infection with the relatively benign cowpox virus induced cross-reactive immunity that protected against the far more lethal smallpox virus (Riedel, 2005). Furthermore, understanding co-infections will help instruct treatment patterns as while in a synergistic relationship treating one of the co-infections may benefit the animal, in an antagonistic situation by medicating one of the infections this may inadvertently allow a previously suppressed pathogen to flourish and impact the trajectory of the disease (Thumbi et al., 2014). An example of this is seen in gut microbiomes, when the use of antibiotics disrupts the balance of bacteria, as seen in *Clostridioides difficile (C. difficile)* where antibiotic use will lead to dysbiosis enabling for the proliferation of *C. difficile* which can lead to cases of diarrhea and mild belly cramping or in severe cases kidney failure (Mada & Alam, 2025).

Some of the most common haemopathogens seen co-circulating calves suffering from ECF are species of *Babesia*, *Anaplasma* and *Ehrlichia.* Babesiosis, is transmitted via ticks and primarily goes on to impact a host's erythrocytes. The disease is characterised by fever, anaemia, hepatosplenomegaly and jaundice, which can result in significant mortality if untreated (Zimmer & Simonsen, 2025). *Anaplasmosis phagocytophilum* and related *Anaplasma* species cause *Anaplasmosis*, a disease transmitted by tick bites. The disease characteristics are fever, headache, chills and muscle aches (CDC, 2025; Tabor, 2022). Lastly *Ehrlichia* species like *Ehrlichia ruminantium* is the causative agent of heartwater, a form of Ehrlichiosis. A tick borne disease characterised by the accumulation of fluid in organs such as the heart and lungs. Co-infections like these may complicate the diagnosis and influence the host's immune response to ECF altering the outcome of disease (*Ehrlichiosis*, 2018; Meyer et al., 2023). Therefore it is important to consider the epidemiological roles of these additional haemopathogens.

Historically, haemopathogens were diagnosed using light microscopy, which, while useful, has several limitations. Microscopy is relatively insensitive, requires skilled personnel, and often struggles to detect co-infections, especially when pathogen densities are low (Vazquez-Pertejo & Bush, 2025). More recent laboratory processes do enable easier diagnoses of these infections through various techniques such as multiplex qPCR, which allows for detection of multiple infections prior to the body planting an immune response, providing a more unbiased estimate of infection prevalence and facilitating the study of co-infections (Porcelli et al., 2024). Other more recently used methods are also showing great promise such as the analysis of blood lipids and protein profiles. This method has the potential for more tailored diagnoses techniques. Laboratory diagnostic methods, however, do still have limitations, for example, although multiplex qPCR is a powerful diagnostic tool it still entails a complex primer design, the need for specialized equipment and trained individuals. This may result in quite a costly method which is difficult to implement in resource-limited settings (Lei et al., 2021).

Treatments and control of co-infections present particular challenges; however, failure to address them threatens increased disease severity, higher mortality, and significant economic losses for farmers. There is no single definitive treatment for co-infections, and addressing

multiple pathogens simultaneously often increases both the cost and complexity of therapy. In many cases, existing strategies are insufficient. For example, vaccination alone may be insufficient to manage co-infections effectively, one reason for this is the limitation of vaccination programs that typically target single pathogens. Furthermore when faced with the case of haemopathogens such as *Theileria, Babesia,* or *Anaplasma*, these vaccines often provide limited protection due to factors like strain diversity, immune evasion, reliance on cell-mediated immunity (Alzan et al., 2024; Barbet et al., 2001) . There are also practical constraints to administering vaccines in resource-limited settings with frequent inabilities from supply chains to provide the necessary availability of vaccines at the administration points (Boeck et al., 2022). Vector control, on the other hand, offers an approach that is able to simultaneously reduce the incidence of multiple different haemopathogens by targeting the common vectors. Acaricides for example are pesticides designed to kill ticks and mites. While effective when this treatment was first implemented, their efficacy has increasingly been undermined since, due to the emergence of acaricide-resistant tick populations, making long-term control challenging (Obaid et al., 2022). Furthermore control measures like this are usually expensive and not easily maintainable as constant re-application is necessary. In order to address these broader challenges in the control and treatment of co-infections, improved diagnostics would be necessary, with affordable and easily attainable tools being at the forefront of the requirements. This could be further supported by educating farmers to identify disease signs and on biosecurity in general. More investment into research and surveillance would help to support the management of co-infections (Graf et al., 2004).

1.3 East Coast fever

One of the most important causes of mortality in cattle in East, Central and Southern Africa is ECF which is responsible for an estimated minimum 1.1 million cattle deaths per year, and a loss of over 300 million US dollars per year (*East Coast Fever*, 2015; Toye et al., 2020; Wragg et al., 2022). ECF is a disease caused by *T. parva*, an intracellular apicomplexan protozoan pathogen, and transmitted by *Rhipicephalus appendiculatus*, a tick vector. Its impact on cattle health and productivity can be better understood by examining its pathogenesis and disease dynamics (Fry et al., 2016).

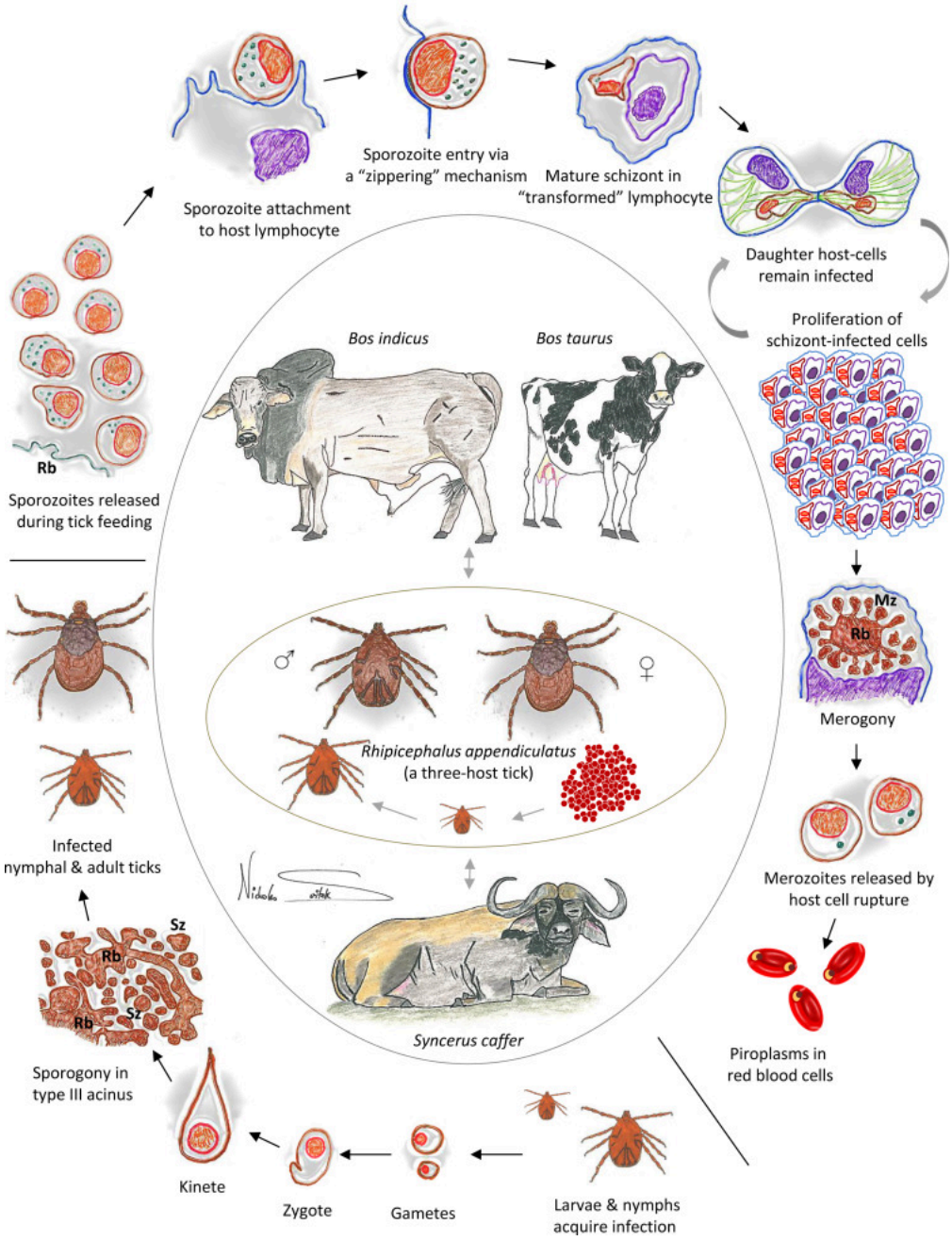


Figure 1: Life-cycle of Theileria parva *illustrating life cycle stages of the pathogen as when it is in both the mammalian and tick host (Nene et al., 2016).*

1.3.1 Parasite life cycle and transmission

Figure 1 details the complex life cycle of *T. parva*. This cycle begins when a larval or nymphal tick feeds on an infected host and ingests the piroplasms found within the bovine erythrocytes (J. A. W. Coetzer, 2005). In the tick's gut, the piroplasms will undergo sexual developments and differentiate into globular macrogametes and thread-like microgametes. The fusion of these two gametes is the formation of zygotes who are responsible for invading the gut epithelium (J. A.

W. Coetzer, 2005). During the tick's immature stages it will undergo moulting, when this is happening the pathogen will continue developing. The zygotes will produce motile club-shaped kinetes that circulate the haemolymph and invade epithelial cells of the salivary glands and develop to form large syncytial sporoblasts from which thousands of elongated sporozoites are produced, ready to be transmitted when the tick next feeds (J. A. W. Coetzer, 2005).

During the tick's next blood meal, the sporozoites are released into the tick's saliva and invade the bovine bloodstream. Then, they are able to rapidly invade lymphocytes and by around 3 days post-infection are differentiating into multinucleate schizonts. The infected lymphocytes will transform into lymphoblasts which allows for simultaneous host cell proliferation and schizont replication (Fry et al., 2016). The host will then experience an exponential increase in pathogen cell dissemination throughout the lymphoid tissue and subsequently into the non-lymphoid organs. This large-scale dissemination is facilitated by pathogen-driven production of a T-cell growth factor which is functionally similar to interleukin-2 (J. A. W. Coetzer, 2005). The pathogenicity experienced by the host at this stage does not stem from sporozoite infections of T and B cells but instead as a result of uncontrolled expansion of infected T-cell populations (J. A. W. Coetzer, 2005; Nene et al., 2016).

Initially this lymphoblast proliferation will be localised at the site of inoculation however through the next 5 days will spread to the draining lymph nodes, this coincides with the onset of fever. Over time these macro-schizonts will transition into micro-schizonts, which will re-enter erythrocytes as piroplasms, completing the cycle (J. A. W. Coetzer, 2005). The massive expansion and subsequent destruction of lymphocytes will lead to severe immunopathology, and infected animals become highly susceptible to secondary infections, contributing to high mortality (J. A. W. Coetzer, 2005).

1.3.2 Pathology and Clinical Manifestations

In advanced stages of ECF, mortality is primarily driven by respiratory failure resulting from pulmonary oedema (*Understanding East Coast Fever–Tickborne Diseases in Uganda*, 2025). This happens when infected lymphocytes recruit and activate macrophages, a process exacerbated by the pathogen driven dysregulation. This accumulation causes macrophage mediated vascular damage in small blood vessels, particularly in the lungs. In damaging the endothelial liming the vascular permeability will increase and leaky vessels will allow for fluid to escape into surrounding tissues such as the lungs. The resulting pulmonary oedema severely impairs lung function, contributing to respiratory failure. Another key clinical outcome of this life cycle is lymphoproliferative syndrome, a condition whose incubation will generally last 15 days, these syndromes are responsible for signs such as enlarged lymph glands, listlessness and peripheral lymphadenopathy, fever, anorexia and respiratory distress (Fry et al., 2016; J. A. W. Coetzer, 2005; Lacasta et al., 2018). In terms of productivity of the cattle these will impact milk production, reproductivity and can ultimately cause the death of the animal (Gachohi et al., 2012; Nanteza et al., 2023; Nicholson et al., 2019).

Neurological signs have also been reported in cattle due to the blockage of cerebral veins by clusters of pathogenised cells, effectively acting as emboli and impairing blood flow (Fry et al.,

2016). An estimated 95% of cattle will recover, with immunity primarily mediated through cell mediated immune response. A Cytotoxic T-lymphocyte (CTL) is able to recognize *Theileria* antigens on the surface of the infected lymphocytes meaning the CTL's cytotoxic activity is restricted to that pathogen. When re-exposed memory T-cells will be activated re-stimulating the release of CTL cells. The circulating antibodies against the schizonts will form the basis of diagnostic serological tests (J. A. W. Coetzer, 2005).

1.3.3 Diagnosis and surveillance

The definitive diagnosis of this disease is nuanced as it relies on clinical approaches making ECF control a challenge. Diagnosis is possible through blood smear microscopy to detect piroplasms, serological tests such as ELISA, and PCR-based methods to identify *T. parva* DNA in blood (Nanteza et al., 2023). In smallholder settings however, access to laboratory diagnosis is restricted and farmers may instead rely on clinical signs such as enlarged lymph nodes, fever or anorexia (Irvin & Mwamachi, 1983; J. A. W. Coetzer, 2005). This is a challenge as these signs are non-specific and will overlap with those of other livestock diseases. This would complicate the administration of correct medication to the animal in a timely manner, threatening their health. If incorrect medication is administered this would incur the farmer further costs. Tick and therefore vector surveillance are also an important component of disease monitoring (Irvin & Mwamachi, 1983). Mapping tick distributions is a promising tool to support decision making for control strategies which uses modern technologies to allow for identification of higher risk areas of ECF outbreaks. Through understanding tick population dynamics and environmental factors influencing tick spread, and the identification of areas most vulnerable to outbreak, targeted interventions can be implemented in these areas such as vaccination strategies or acaricide campaigns. This will ultimately aid in control and reduce control costs as it will address outbreaks before they become too severe (Irvin & Mwamachi, 1983).

1.3.4 Vector ecology and control

The *R. appendiculatus* vector belongs to the Ixodidae family and can be distinguished by its red/brown coloration, its ornate scutum as well as the long mouth parts which are adapted to be able to pierce the skin of the host to allow feeding (Klompen et al., 2001). Nymph stages of this vector prefer to feed on the ears, head and legs of the cattle whereas adult stages are more commonly found on the pinnae of the ears and head (Klompen et al., 2001).

The ticks' life cycle is composed of 4 stages: egg, larva, nymph and adult. The life-cycle begins when the tick eggs are laid in the soil and hatch into larvae which then will climb vegetation in search of a host on which to feed. After feeding the larvae will detach from the host and moult into the nymph stage. This subsequent step in the life-cycle sees the nymph search for a new host on which to feed and detach again. This time when it moults it will enter the adult stage of its life. An adult tick will find a final host for a final feed of blood, this feeding will provide the nutrients necessary for female ticks to produce thousands of eggs. Once engorged the female will detach and lay these eggs. This completes the three-host life cycle which can take between several months and years to complete (Bouchard et al., 2019; Onyiche & MacLeod, 2023).

R. appendiculatus are able to thrive in warm and humid environments. Vegetation cover gives the ticks shelter as well as a microclimate suitable enough for egg hatching and larval survival. East Africa therefore offers a perfect climate with moderate to high rainfall and also temperature of between 20°C and 30°C (Bouchard et al., 2019; Onyiche & MacLeod, 2023). These environmental factors also shape the distribution of T. parva, the parasite responsible for ECF, which predominantly affects cattle across central and southern Africa, especially in Kenya, Uganda, and Tanzania (Allan & Peters, 2021; Byaruhanga et al., 2017). Its abundance in these regions is influenced by environmental conditions such as soil moisture, temperature, habitat suitability and host availability. While broader climate factors also play a critical role in shaping the epidemiology of ECF. A study showed a negative association between monthly rainfall and mortality in Boran cattle associated with ECF, this was representative of a 2-month lag led by tick ecology and environmental and physiological factors (Chepkwony et al., 2020).

ECF epidemiology is also linked to livestock movement patterns, for instance when livestock are traded or moved for grazing purposes (Gachohi et al., 2012). Within affected herds, these ecological and epidemiological dynamics result in different patterns of disease expression, within herds the disease can have endemic stability, endemic instability or be an epidemic (Wragg et al., 2022). When in epidemic form this will likely be because the disease has been introduced to areas that were previously free of ECF. In contrast, in endemic regions of Kenya, many cattle persist as carrier animals, this means that cattle will either be clinically ill, persistent carriers or have recently recovered (Wragg et al., 2022).

Tick vectors make the control of ECF challenging for numerous reasons; the misuse of acaricides to combat ticks have made certain populations of the tick resistant. Furthermore, the constant use of these chemicals is also a risk to ecosystems as they will impact non-target organisms raising sustainability concerns. Climate change also complicates tick control, as rising temperatures and shifting rainfalls are contributing to expanding tick habitats into areas that were previously unsuitable for their survival (Bouchard et al., 2019). The estimated distribution of ECF is shown in Figure 2.

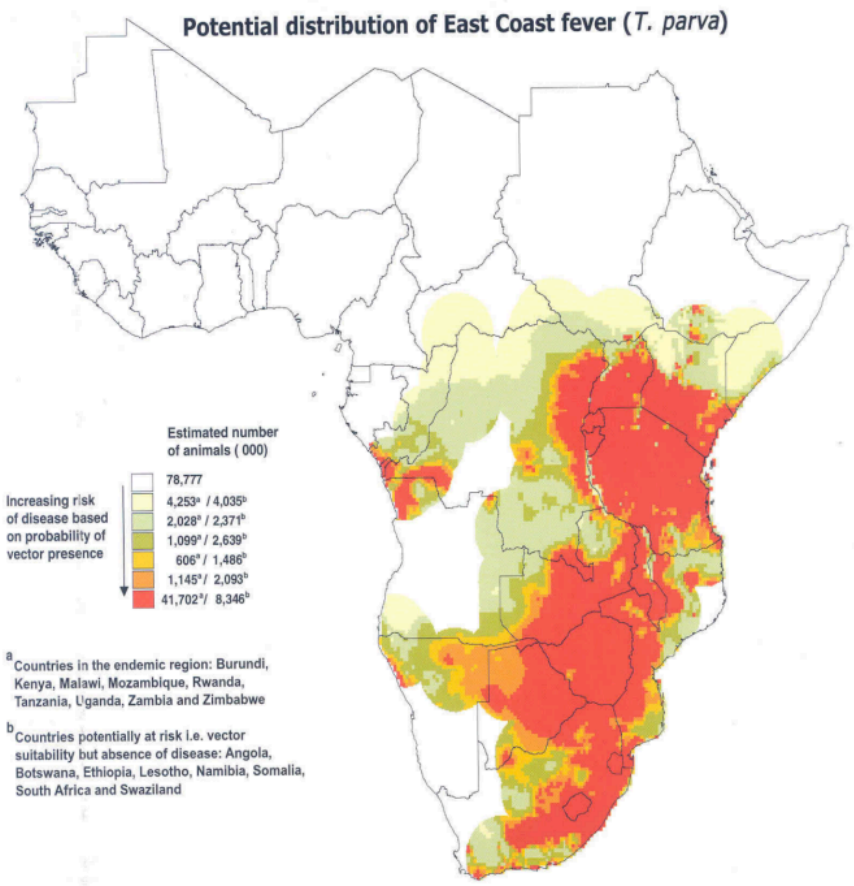


Figure 2: Estimated distribution of Theileria parva in animals across Africa *(de Villiers, n.d.)*.

1.3.5 Current Control Strategies

In the past, control of ECF has relied on control of the tick vector, these strategies include the use of acaricides, chemical compounds used to inhibit a ticks ability to successfully attach and feed on cattle. These compounds fall into several families (such as organophosphates, amidines, carbamates, and synthetic pyrethroids). Most acaricides operate as neurotoxins and work by disrupting a ticks nervous system; one example of this can be found in carbamates which inhibit acetylcholinesterase, this causes paralysis and death in ticks (Nwanade et al., 2022). The application of this chemical includes by dipping, spraying or pouring of the chemical over the impacted areas. Through the reduction in tick populations and movement the impact of the disease can be mitigated, this over-reliance however has led to the emergence of acaricide-resistant tick populations in some areas. This is driven by the frequent and improper use of the chemical such as both over-dosing and under-dosing (Obaid et al., 2022). Recent studies from Kenya, showed up to 60% of studied ticks from *R. appendiculatus* and *R. decoloratus* were ‘super resistant’ which meant showing 0% mortality in response to the acaricides cypermethrin and deltamethrin (Githaka et al., 2022).

Another form of control is immunization of animals through the Muguga cocktail vaccine, this live vaccine contains 3 species of *T. parva,* those being Muguga, Kiambu 5, and Serengeti transformed, which allows cross-protective immunity. While the vaccine procedure does have a narrow therapeutic index, meaning the margin between effective dose and harmful dose is small, the animals then acquire life-long immunity to ECF (Allan & Peters, 2021; Tamargo et al., 2015). Although available for almost 4 decades and with over 1.5 million doses delivered by 2016, availability has been inconsistent (Surve et al., 2023). This is due to manufacturing complexities, as there were doubts as to whether consistent batches were possible in production. Epidemiological concerns also arose as the 3 species used cannot confer protection against all species throughout impacted areas. Lastly the requirements necessary for the production and storage of the vaccine are not widely achievable through infected areas as firstly the vaccine has not been obtained by a private commercial organisation, and the storage requirements such as liquid nitrogen and trained personnel to administer vaccinations have not been met (Allan & Peters, 2021). The vaccine's efficacy also shows regional variation. For instance, cattle populations in close proximity to buffalo populations acquire less effective protection. This is as a result of genetically diverse *T. parva* species hosted by buffalos that when these buffalo-derived parasites spill over to cattle through vectors, cause Corridor disease, a severe form of ECF that current vaccines do not offer protection from (Allan & Peters, 2021; Cook et al., 2021).

1.3.6 The Role of Co-infections

The impact of co-infections on ECF progression should not be underestimated and could confer significant mortality differences (Onyiche & MacLeod, 2023). One such co-infection that should be further investigated is the differential pathogenicity of *Theileria* species in early calfhood and how prior exposure to less pathogenic species, specifically the less pathogenic species *Theileria velifera* and *Theileria mutans,* may influence disease outcomes (Chaisi et al., 2013; Woolhouse et al., 2015). These *Theileria* species cause transient clinical signs but rarely lead to death. Although *T. parva. T. velifer* and *t. Mutans* are distinct species, they are closely related and therefore often share epitopes of conserved antigens. This is especially evident in antigen-encoding genes such as p67 and Tp1-Tp10, which are key targets of CTL cell responses (Woolhouse et al., 2015). Therefore early exposure to these milder forms of *Theileria* will act as a primer for the hosts immune system conferring cross-protection in order to reduce disease severity long term. However, studies have indicated a large amount of variance in diversity between *theileria* antigens recognized by CTL cells and so this may not be a completely reliable form of protection (Sitt et al., 2018). Other ways this prior exposure may offer modulatory effect is through the stimulation of the innate immune pathway, a heightened immune vigilance state will act faster when a *T. parva* infection occurs (Morrison, 1984). This is a highly relevant topic in endemic regions where multiple *Theileria* species co-circulate, and so the relative virulence of each strain should be understood and used to elucidate further on the hypothesis that the well-known immunomodulatory effects of *T. parva* are moderated by prior exposure to less pathogenic species such as T. mutans or T. velifera, thereby reducing the severity of infection. If these protective effects are confirmed, vaccination approaches or controlled exposure strategies could be developed as potential prophylactic tools (Woolhouse et al., 2015).

In addition to Theileria co-infections, helminths such as Haemonchus contortus also play an important role in shaping disease outcomes. Haemonchus is a highly pathogenic blood-feeding nematode and is responsible for inducing anemia and hypoproteinemia (Flay et al., 2022). This can be catastrophic for the calf as it will weaken them during the critical window of ECF exposure. Moreover, helminth infections are known to skew host immunity toward Th2-type responses, potentially impairing the cytotoxic T cell activity that is essential for controlling T. parva infections (Thumbi et al., 2014; Moreau et al., 2010). This means that calves burdened with Haemonchus may not only be physiologically compromised but also immunologically less able to mount effective responses, increasing the risk of severe disease. This is an especially big concern in endemic regions where both parasites co-circulate as controlling these interactions could significantly alter patterns of morbidity and mortality (Flay et al., 2022; Thumbi et al., 2014; Moreau et al., 2010).

1.4 Key factors of interest for this study

To effectively control ECF, it is essential to understand the key factors that influence infection risk and mortality in calves. This study investigates several of these factors, including host genetics, co-infections with other haemopathogens, pathogen dynamics over time, the influence of calf sex and environmental influences such as sublocations. By identifying factors affecting disease outcomes, we aim to contribute insights that could support future interventions, inform veterinary policy, and guide more targeted education and management strategies.

One hypothesis is that variability in disease susceptibility in cattle is genetically based. Taurine breeds are found to be more susceptible to infections than Indicine breeds such as Zebu cattle. A recent study investigating host genetic factors and their impact on ECF progression, showed evidence of genetic tolerance in a cohort of Boran cattle, where a 6 Mb genomic region of their bovine chromosome 15 was found to be statistically associated with cattle survival after *T. parva* exposure (Alcaraz-López et al., 2021; Lee et al., 2024). This same region has previously been implicated in heritable tolerance to buffalo-derived *T. parva* infection, the more virulent form of the pathogen responsible for ECF in an extended cattle pedigree (Wragg et al., 2022). The protective effect is linked to a single nucleotide polymorphism (SNP) found on chromosome 15 in a specific gene called the *FAF1* paralog. The *FAF1* paralog is a gene related to the Fas-associated factor 1 whose role is to regulate cell death via the Fas-induced apoptosis pathway, a mechanism which *T. parva* exploits to promote proliferation of infected lymphocytes. The SNP changes a cytosine (C) to thymine (T), converting what would be an arginine in the wild-type gene into a premature stop codon. The three possible genotypes in this allele therefore are as follows: CC (wild-type-homozygote), CT (heterozygous) or TT (variant homozygote) (Wragg et al., 2022). The study found animals carrying the homozygous alternate allele showed complete resistance with 0% of animals succumbing to *T. parva* infections compared to 53% that were homozygous for the wild-type allele (Wragg et al., 2022).

Previous studies have described protection to come from this phenotype through a delayed and improved parasitosis and febrile reaction in the cattle, consistent with a modified apoptotic

response (Miyunga et al., 2025). Furthermore, evidence for protection with TT allele dosage has been reported, this is consistent with recessive or dosage-dependent genetic resistance. This is evidenced in prior studies with calves carrying a heterozygous allele CT displaying less clinical signs and surviving for several more days than those not carrying a copy of the variant (Wragg et al., 2022). In vitro studies also showed that lymphocytes in these mutated phenotype cattle infected by *T. parva* would obtain less proliferation (Miyunga et al., 2025). This raises potential for marker-assisted selection of tolerant cattle (Wragg et al., 2022).

Calf sex is a potential influencer in susceptibility to ECF through shaping immune responses and mortality outcomes. Evidence for sex-based differences in immune response and disease susceptibility/progression is widely seen. While not a primary risk factor for tick exposure (when mixed-sex herds are managed similarly), their differences in physiology, hormonal status and management factors (e.g., preferential treatment, nutritional access, or workload) may influence the time to infection and mortality risk (Barger, 1993; Broughan et al., 2016; Klein & Flanagan, 2016).

Investigating the timing of infection, specifically early infection (before 25 weeks of age) versus later is critical for understanding disease dynamics (Ngetich et al., 2025). Early-life infections can have fundamentally different implications than those acquired after the neonatal period. Calves under 25 weeks old are in a transitional immunological window, relying heavily on maternal antibodies (via colostrum) while their own immune systems are still developing (Chase et al., 2008). Relying on this passive immunity can offer a calf protection from the disease however simultaneously it disallows a calves ability to develop its own immunity to the invading pathogen. If maternal protection is also too inadequate the calves immune system may be too immature to mount an effective response. The selection of the 25-week cutoff is supported by evidence from studies on the development of active immunity in calves. In the first four weeks of life, immunomodulatory mechanisms suppress Th1 responses, which are essential for long-term immune memory, while promoting Th2 responses associated with short-term immunity (Chase et al., 2008). During this neonatal period, calves also have relatively few dendritic cells, limiting their capacity for antigen presentation and the establishment of adaptive immunity. By eight weeks of age, however, many key immune components have matured; notably, circulating natural killer cells increase to around 10% of total lymphocytes, marking a transition toward a more functional and responsive immune system. Furthermore, B-cells circulation will rise to 20% of lymphocytes circulating the body (Chase et al., 2008). Therefore the 25-week mark represents a point beyond the neonatal period at which early immune developmental processes no longer impact the course of ongoing infections. This hypothesis will be used to investigate survival models based on time of infection in a calves life. Infection trends based on age of calf at infection is worth investigation as if early infections are shown to be more dangerous, or if early non-lethal exposure improves outcomes it may influence when and how interventions like vaccination, tick control, or breeding for tolerance should be applied (Chase et al., 2008).

This study will therefore examine the epidemiology and impact of a number of risk factors on ECF related mortality, this will be done through the incorporation of host genetic, haemobiome data along with all the environmental and covariate information per calf available in the IDEAL

databases. This will be used to statistically evaluate the risk factors and predictors associated with disease outcomes and create a foundation from which to begin to inform more advanced and specific control strategies.

2. Materials and Methods

2.1 Study design

This study is a re-analysis of biobanked samples collected as part of the longitudinal IDEAL (Infectious Diseases of East African Livestock) project, which ran from 2007 to 2009 using new molecular typing tools (Yalcindag et al., 2024). The IDEAL project aimed to elucidate the relationship between co-infections in a cohort of cattle and their resulting impacts on mortality and clinical infections. This present study, specifically, will focus on the role of haemopathogens, co-infections, and host genetics in shaping mortality through the use of temporal and survival analyses techniques (Callaby et al., 2020).

2.2 Study population and sampling criteria

2.2.1 Sampling criteria

The IDEAL study was conducted in an area of western Kenya where 20 sublocations (administrative units comparable to parishes in the UK) were defined from which to recruit the calves. In order to select these sublocations in an unbiased manner the original study design drew a 45 km semi-circle around Busia, and all sublocations falling within this area were classified according to agroecological zones (AEZs) (Figure 3) (de Clare Bronsvoort et al., 2013). From this, 20 sublocations were then selected by stratified random sampling across AEZs to ensure a representative spread of environmental and management conditions. The predominant breed of cattle kept amongst the farmers sampled in this region was Zebu cattle with a smaller percentage of 3.1% keeping Zebu crosses (de Clare Bronsvoort et al., 2013; Callaby et al., 2020).

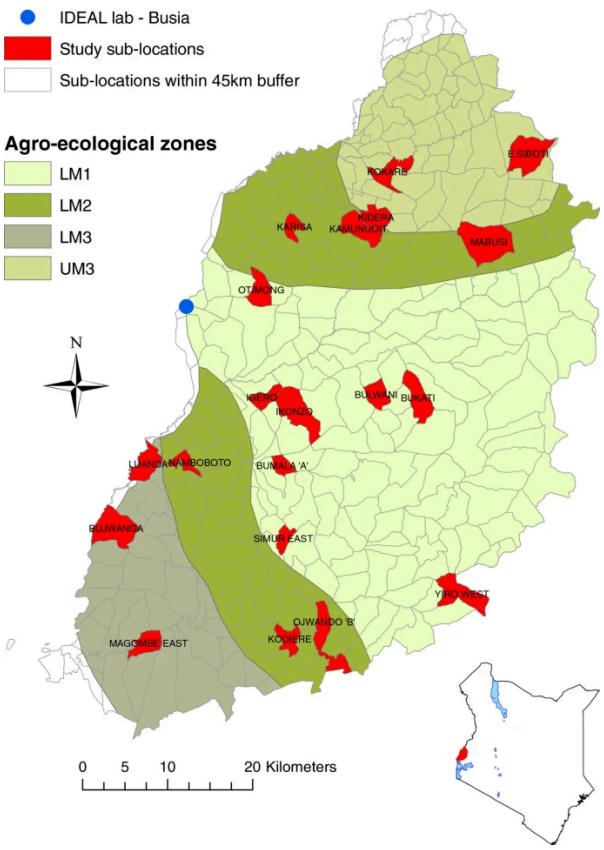


Figure 3: A map showing the study areas chosen for this study in western Kenya. *Agro-ecological zones shown in 4 shades of green, sublocations shown in Red (de Clare Bronsvoort et al., 2013).*

The specific production system investigated in this project was livestock small holdings kept for meat production purposes. A total of 548 calves were recruited as part of this epidemiological study, in order to ensure consistent exposure and minimize variability in maternal immunity and management history calves had to have been recruited between 3 to 7 days of birth and born to a dam living at the farm for at least a year. By including calves between 3 to 7 days of age the study was able to carry out uniform longitudinal follow-up from the neonatal period (de Clare Bronsvoort et al., 2013). However only one calf per farm could be included to prevent the possibility of household-level-clustering and control for shared environmental and management factors. If herds were practicing stall feeding, a management strategy that reduces the contact between tick infested areas and cattle, they were excluded due to the risk of homogeneity in tick exposure and if the calf had been conceived as a result of artificial insemination they were excluded to avoid the possibility of genetic confounding as a result of sire variability. Efforts were made to include calves with more European genetics as evidence of higher ECF susceptibility have been shown in those genotypes (Murray et al., 2013; de Clare Bronsvoort et al., 2013; Callaby et al., 2020).

Calves in the study had to remain on their birth farms and managed under the typical conditions of the farm with minimal or no veterinary interventions. If any kind of anthelmintic or antibiotic

was used throughout the study the data from affected calves at the time of treatment was censored. No preventive vaccines and treatments were used during this time, this was done to ensure the data only reflected natural pathogen exposure and the development of immune responses to endemic infections ensured the data only reflected natural dynamics of pathogen exposure and built immune responses in the absence of external controls. If welfare concerns required the intervention of veterinary services, animals were censored from the dataset from the point of intervention forward (de Clare Bronsvoort et al., 2013; Callaby et al., 2020).

2.2.2 Ethical approval

The original study was approved by the University of Edinburgh Ethics Committee (reference number OS 03-06) as well as the Institute Animal Care and Use Committee of the International Livestock Research Institute, Nairobi. Prior to recruitment of animals into the study, informed consent from the farmers involved in the study was obtained in their native language (Callaby et al., 2015).

2.3 Sampling strategy

2.3.1 Routine sample visits

Routine visits were conducted at five week intervals from birth until up to one year of age or death. At time of recruitment and every time a clinical illness was reported a complete clinical examination was undertaken in order to screen for pathogens through the use of haematological and immunological tools. These samples were blood smears, whole blood, serum samples, and fecal samples (de Clare Bronsvoort et al., 2013; Woolhouse et al., 2015; Callaby et al., 2020).

2.3.2 Clinical visits

All clinical illnesses required reporting by the calves' owners or a resident local animal health assistant, and within 24 hours a detailed examination was conducted in a standardized manner. Clinical ECF was determined through clinical signs and laboratory detection of *T. parva.* Examples of clinical signs were rectal temperature of above 40^{o}C as well as enlarged lymph nodes. Most times the parotid lymph node was the one used however, enlarged precrural and suprascapular lymph nodes were used for diagnosis in one case (de Clare Bronsvoort et al., 2013; Woolhouse et al., 2015; Callaby et al., 2020).

2.3.3 Postmortem visits

When an animal died as part of the study whether through disease or euthanasia a complete post-mortem would then be conducted (a standard body system-by-body system veterinary postmortem routine). The etiological cause of death in each case was determined by the study team through parasitological and histological reviews of samples taken postmortem as well as gross-lesions observed on the animal's body (de Clare Bronsvoort et al., 2013; Woolhouse et al., 2015; Callaby et al., 2020).

2.3.4 Bias

Seasonal and spatial variation is an important consideration for this sampling and so recruitment of calves were staggered over 2 years to combat this. In order to avoid bias, clinical episodes and deaths were recorded without any knowledge of the calves' infection status. In order to collect the calves blood samples and fecal samples standard techniques were used, all interaction with the animal was conducted by either a professional animal health assistant or a qualified veterinary surgeon. Veterinary surgeons were further available for examination of the calves during sickness and clinical episodes. If a calf was found to be in severe distress as a result of illness, the calf would be euthanized via injection with sodium pentobarbital. Recall bias was mitigated through thorough documentation at each visit. However, it is important to note that Infection timing was inferred from visit data and not continuous sampling so the true exposure time of calves to haemopathogens is approximate (de Clare Bronsvoort et al., 2013; Woolhouse et al., 2015; Callaby et al., 2020).

2.4 Data Sources

2.4.1 IDEAL databases

The IDEAL database was used as a data resource, found online at http://data.ctlgh.org/ideal/ , the database consists of 8 tables detailing all the data collected during the IDEAL project. This consists of (de Clare Bronsvoort et al., 2013; Callaby et al., 2020):

- Farm Information includes the specific farms characteristics and the animals kept there as well as its management and husbandry practice.
- Calf information is a table encompassing data from every calf on each of their visits, this includes health assessments, body conditions scores, weight, girth measurement, rectal temperatures, presence of ectoparasites, herd movements, illness within the herd as well as calf phenotypes.
- Dam information, is a table with information from every dam on each visit, this includes health assessments, body condition scores, girth measurements, california mastitis results, udder disorders.
- Test information, a table which shows all diagnostic tests. This will include the haematological profiles carried out on blood, and other samples collected from calves and dams at each visit.
- Clinical information, this will hold a record of all clinical episodes seen in each calf and which body part was affected, the type of disorder seen and the severity of the lesion.
- Post-mortem information, this provides the detailed reports of the death and causes of death of the 88 calves that died.
- Sample information, this records the sample collection details and holds a record of which sample in the biobank is related to which visit.
- A follow-up questionnaire was also conducted a year after the project's completion in order to track how the subjects of the project progressed.

2.4.2 Data cleaning

All data processing, statistical analyses, and visualizations were performed usingR Statistical Software (4.4.0) (R Core Team, 2024). The following R packages were used:

- dplyr (Wickham, 2023)
- ggplot2 (Van den Brand, 2025)
- survival (Therneau, 2024)
- survminer (Kassambara, 2025)
- scales (Wickham, 2025)
- coxphf (Heinze, 2023)
- tidyr (Wickham, 2024)
- viridis (Garnier, 2024)
- ggpubr (Kassambara, 2025)
- ggsci (Xiao, 2024)
- janitor (Firke, 2024)
- complexUpset (Krassowski, 2021)
- gtsummary (Sjoberg, 2025)

Several datasets from the IDEAL database were integrated, cleaned, and transformed to prepare the final dataset used in this analysis.

2.4.4 Linking Sample data from IDEAL database

Sample IDs were matched to IDEAL Visit IDs in the ‘sample information’ dataset from the IDEAL biobank. This allowed integration of Haemobiome results into the calf metadata.

2.4.5 Linking IDEAL Calf information

Data from the IDEAL ‘Calf information’ dataset (e.g., dates of birth, visit dates, health assessments, mortality records) were merged with the MiSeq dataset.

- Date fields were cleaned and reformatted to proper Date types.
- Sample week was derived from Visit ID or if found to be wrong it was instead calculated using date of birth and sample date with this formula.
  $Sample\ week = \frac{Sample\ date - Birth\ date}{7}$
- Records were filtered to exclude virtual re-sampled visits (non-biological repeats).
- The Routine Visit on Dams (VRD) were removed leaving only the Routine Visit on Calves (VRC) and Clinical Visit on Calves (VCC).
- Calves were classified as male or female at birth and this binary classification was used throughout all statistical models.

2.4.7 Linking post mortem and cause of death data information from IDEAL database

Postmortem records and cause of death diagnoses were integrated using calf ID. In addition to the exact cause of death, a simplified classification was created to distinguish “Dead”, “Alive”, or

“Censored” calves, depending on death status and cause. In order to not be considered ‘dead’ in the statistical models the causes had to have aetiologies and pathophysiological mechanisms unrelated to *T. parva* and occurred independently of an ECF infection status. Censoring these cases is an adequate categorization as it acknowledges the calves exiting the study therefore taking them into account in the time-to-event analyses however excludes them from the event of interest. Furthermore, including these unrelated deaths would cause misrepresentation of the mortality risk attributable to ECF.

- Dead were therefore those calves that were classified as dying from ECF in the ‘definitive aetiological cause’ column of the Post-Mortem IDEAL Dataset.
- Alive were those with NA in the ‘definitive aetiological cause’ column of the Post-Mortem IDEAL Dataset.
- Censored were those with any of the following as a cause of death in the ‘definitive aetiological cause’ column of the Post-Mortem IDEAL Dataset: Haemonchosis, Unknown, Foreign body, Actiomyces pyogenes, Trauma, Heartwater, Trypanosomiasis, Turning sickness, Cassava, Mis-mothering, Bacterial pneumonia, Black quarter, Viral pneumonia, Rabies, Arcanobacterium, Babesiosis, Salmonellosis.

2.4.8 Survival Time and Event Definition

Event status was defined as 1 for death and 0 for censored/alive.

- Time to event was calculated as weeks between birth and death or if no death date was provided then date of last follow-up was used.
- For calves with incomplete records or who were alive at study end, time was capped at 51 weeks.
- First detection of a pathogen was defined as the earliest sample week in which the strain/Haemopathogen of interest was detected in NGS (MiSeq) reads in a calf.
- For comparison of the impact of early vs late infections, an early pathogen infection was defined as the first detection of the specific pathogen of interest in the calves MiSeq data before 25 weeks of that calves life and a late infection was defined as first detection of specific pathogen of interest after 25 weeks.

2.4.9 Linking of sublocations from IDEAL database

Sublocations data were added by linking calf IDs to sublocation identifiers from the IDEAL ‘farm information’ database.

2.4.10 Incorporation of Exposure order

Exposure order was integrated into the R Statistical Software script by assigning each calf into one of 3 categories and displaying the result in a column ‘infection_order’. The 3 categories are *No exposure (3 species), T. mutans/T. velifera First, T. parva First.* Classification for the analysis split the calves into three categories,

- *No exposure (3 species)* refers to those calves with no read counts for any of the 3 *Theileria* species *T. mutans, T. parva and T. velifera* throughout a calves life/results in the study.

- *T. parva First refers to* calves that first detected a *T. parva* infection before, or concurrently with either *T. velifera* or *T. mutans.*
- *T. mutans/T. velifera First refers to c*alves that first detected *T. mutans* or *T. velifera* before the detection of a *T. parva* infection, including cases where both species were present simultaneously without prior *T. parva* exposure.

2.4.11 Incorporation of genetic database

The calves within the IDEAL population were genotyped using the 50K Illumina® BovineSNP50 beadchip v. 1. This was made up of 55,777 SNPs before quality control (de Clare Bronsvoort et al., 2013). Quality control procedures before analysis consists of GenABEL in R. SNPs with a minor allele frequency of below 1% or a call rate of below 90% were not included. To not encounter sample misidentification, an identity-by-state threshold of 90% was used with a likelihood ratio cut -off of 1000:1. 9 calves and 13,856 SNPs were removed from this dataset after following this quality control (Murray et al., 2013; Callaby et al., 2020)..

The genetic database was incorporated into the R Statistical Software script using calf ID. Although originally 548 calf genotypes are available, 2 calves were removed due to lack of their calf ID presence in any other dataset and so no other available information with which to confer statistical hypothesis. Calves were classified by genotype as CC (homozygous reference), CT (heterozygous), or TT (homozygous variant).

2.5 Laboratory procedures

2.5.1 Molecular diagnostics: Deep amplicon Pathogen Sequencing Data

High-throughput amplicon sequencing data were obtained from the haemobiome tool dataset, as previously described by Yalcindag et al. (2024). In this approach data from multiple species of *Anaplasma, Ehrlichia, Babesia,* and *Theileria* obtained from genus-specific PCRs were extracted from a hamobiome tool dataset. The haemobiome is a high-throughput approach designed for simultaneous detection of multiple haemopathogen genera in cattle. Genomic DNA was extracted from preserved calf blood and genus-specific PCR primers were designed based on DNA sequences (Yalcindag et al., 2024). The primers amplified the V4 region of the 18S rRNA gene for *Theileria* and *Babesia* and variable regions of the 16S rRNA gene for *Anaplasma* and *Ehrlichia*. A two-step PCR strategy was employed to generate these amplicons with unique identifiers. In the first round, genus-specific loci within the 16/18s rDNA regions were amplified, while the second round added Nextera II multiplex identifier tags (MID) and sequencing primers. The products were subsequently purified using Qiagen QIAquick gel extraction kit (Qiagen) and AMPure XP magnetic beads (1X) (Yalcindag et al., 2024). The pools were the quantified via Qubit (Invitrogen) followed by the sequencing of 70 µL of purified pool on an Illumina MiSeq platform using a 500-cycle paired-end reagent kit (MiSeq Reagent Kits v2, 2 × 250 bp paired-end reads) with 10% PhiX Control v3 (Yalcindag et al., 2024).

2.5.2 Bioinformatic analysis

As described in Yalcindag et al. (2024), a custom-designed bioinformatics pipeline was used to deconvolute data by sample and assign reads to pathogen species. By using multiplexed barcoded primer combinations this method is able to process up to 384 samples simultaneously on a single Illumina MiSeq flow cell (Yalcindag et al., 2024). Raw sequencing reads were quality-filtered (Phred ≥28), merged, and clustered, with singletons, chimeras, and sequencing errors removed. Taxonomic classification was performed using BLAST against the SILVA database, with threshold criteria of ≥99% identity for AnEh and ≥97% for ThBa, and minimum read counts of 1,000 for AnEh and 500 for ThBa to define positive infections (Yalcindag et al., 2024).

Pre-processing for analysis:

- The haemobiome data consisted of 7 datasets, combined into 1 in R
- Missing or NA pathogen read values were set to zero.
- Sample IDs were truncated to a standard format for consistency.
- Repeat samples for the same visit were reduced to one through random selection.

2.6 Statistical analyses

Statistical analyses were carried out using R Statistical Software (4.40) (R Core Team, 2024). A combination of non-parametric tests, regression models and survival analyses was used. Pathogen load differences between groups were assessed using non-parametric tests that compare the median loads between independent groups. Wilcoxon Rank-Sum Tests (Mann–Whitney U Test) were used to compare 2 groups and allow for the analysis of non-normally distributed data with high variance of read counts. A significant difference in this test will imply there is association between pathogen burden and mortality. However the test is not for confounding variables (*Ford, 2017*). Comparison between multiple groups was done by carrying out a Kruskal-Wallis test, a limitation of this test is that it is not able to pinpoint specific group differences and post-hoc testing will need to be done (Kim, 2014). Associations between categorical variables, such as infection timing, were examined using Chi-square. However Chi-square tests have limited power in small sample sizes and so Fisher's exact tests were used where appropriate (Kim, 2017).

2.6.1. Survival analyses

For survival analyses Kaplan-Meier methods were used to estimate the survival probabilities over time. This test is a non-parametric test used to estimate the survival probabilities over time accounting for right-censored data (Dudley et al., 2016; Rich et al., 2010). The variables for this test are the time-to-event measure in weeks, the event itself dead, alive or censored and the various predictor variables such as sex, infection status and genotype. The difference in curves is assessed using a log-rank test where a p-value of below 0.05 will indicate a significant survival difference. This approach provides a descriptive comparison of survival between

groups, but it does not account for multiple covariates simultaneously and therefore only identifies potential associations (Le-Rademacher et al., 2022).

To explore these associations further while adjusting for confounding factors, Cox proportional hazards regression models with Firth's penalization were fitted. A cox proportions hazards model is a semi-parametric regression model which estimates the hazard ratios for death and allows adjustment for multiple covariates (Nagashima & Sato, 2017; Zhang, 2016). This penalized approach reduces bias in small or unbalanced datasets and allows estimation of hazard ratios with corresponding 95% confidence intervals. In this study this was particularly used to address the quasi-complete separation in the genetic data.Statistical significance was set at a p-value of <0.05, although interpretation focused primarily on effect sizes and confidence intervals, as this is more informative for epidemiological inference than sole reliance on p-values. However, this test is less intuitive than a standard cox model and is more sensitive to extreme outliers (Nagashima & Sato, 2017; Zhang, 2016).

These models act as estimators for associations between predictors and mortality risk, this does not mean the models are able to establish causality (Thrusfield, 2018). This analytical approach therefore on top of describing survival patterns is also able to contribute to understanding the factors that might be causally related to ECF mortality apart from simply the presence of infection.

3. Results

3.1 Descriptive analyses

3.1.1 Cohort survival outcomes

The original IDEAL cohort was 548 calves of which new screening data was available for 472 calves. These 472 calves had repeated 5 weekly visits to monitor their health until 51-weeks or until they died or were censored. The disease outcomes related to ECF are given in Table 1, 32 calves died from ECF while 386 survived until week 51. The outcomes over time are shown in Figure 4. The majority of calves however stayed alive till the 51 weeks. Higher rates of death were observed between 5 and 25 weeks of age. dDaths and censored events including calves that died of causes other than ECF occurred over the 51. A breakdown of the number of calf deaths as a result of other conditions is shown in Appendix A Figure 1.

Table 1: Summary statistics of calves survival and mortality associated with East Coast fever. *Summary table showing final numbers for calf survival based on deaths due to East Coast fever. The table shows total calves in the study, total deaths, surviving calves, total censored calves. The median time to event and the mean time to event. 'Event' is representative of death.*

Total calves	Total deaths	Surviving calves	Total censored	Median time to event (Weeks)	Mean time to event (Weeks)
472	32	386	54	51	45.86713

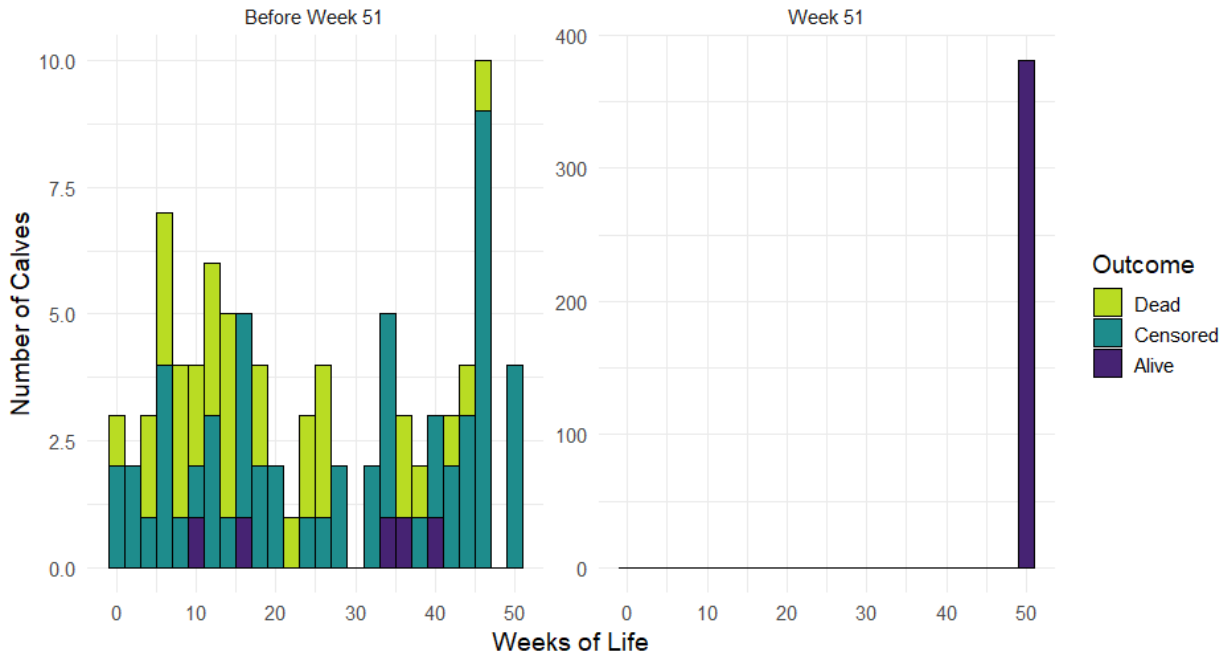


Figure 4: Number of calves that died (light green), were censored (turquoise) or lived (dark blue) based on weeks of age before 51 weeks on the left panel and at 51 weeks on the right panel (shown separately due to scales). *Censored refers to those calves that died or left the study early due to an unrelated to East Coast fever event.*

3.1.2 Pathogen prevalence over time

To further investigate pathogen epidemiology within this cohort, the percentage of calves positive for 4 pathogens genera is shown in Figure 5 and further broken down into the 8 pathogen species in Figure 6. Across the 51 weeks of study high rates of haemopathogen infection were observed. In Figure 5 this can be seen particularly in the *Theileria* pathogen detected in 97.5% of calves, closely followed by *Anaplasma* at 90.9% of calves by 51 weeks. When examining species-level data, Figure 6 shows *Theileria mutans* and *Anaplasma platys-like* had the highest and earliest infection rates, whereas *T. parva* and *T. velifera* showed lower and more gradual infection trajectories.

Further investigation of co-infections in this population and specifically diversity of co-infections throughout a calves life was carried out using a Simpsons diversity index shown in Figure 7. The Simpson index measures diversity, with higher values indicating a greater diversity among pathogens present. During the first week of life, calves exhibited a low diversity (Simpson index 0.2-0.35) indicating infections at this time in the calves' life were more dominated by a smaller subset of haemopathogens. Diversity of pathogens then gradually increased to around 0.6 at week 40 suggesting infections became increasingly broadly distributed across multiple pathogens. A ribbon graph showing the cumulative proportion of calves infected by each of the 5 most prevalent species can be found in Appendix A Figure 2.

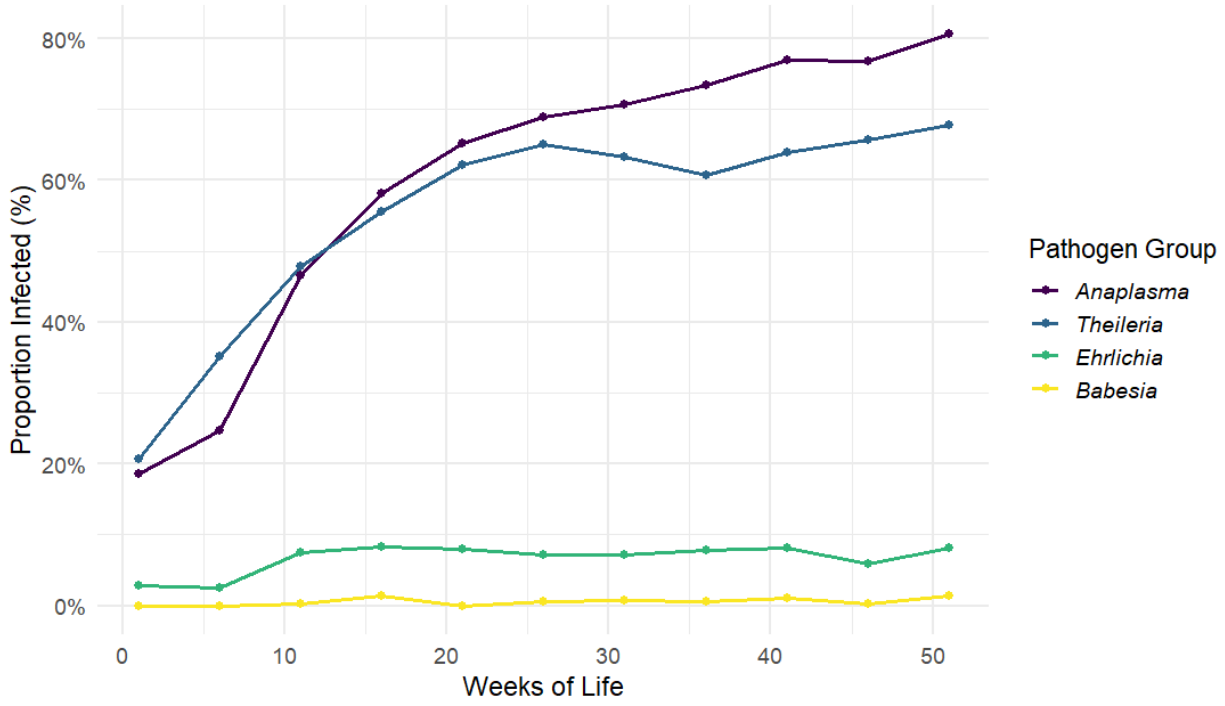


Figure 5: Weekly proportion of calves infected with four haemopathogen genera: Theileria (yellow), Anaplasma (purple), Ehrlichia (green), and Babesia (blue). *Each line represents the infection trend over time for one genus. The proportion infected grew more rapidly for Theileria and Anaplasma with 80% of the calf population being infected with Anaplasma by 51 weeks. Ehrlichia and Babesia remained at below 10% infection of the calf population throughout the 51 weeks.*

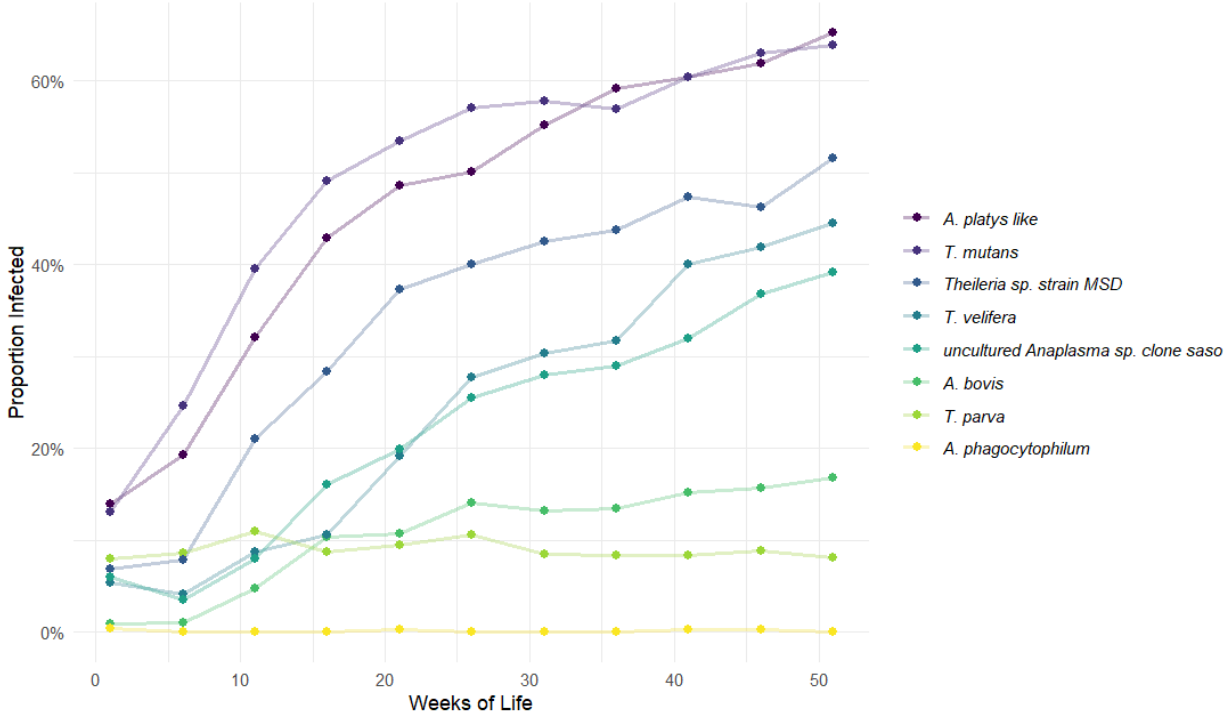


Figure 6: Percentage of calves positive for 8 pathogen species of interest at each sample week. *Weekly trends showed that T. parva burden fluctuated over time. The highest proportion of calves infected are due to infections from A. platys-like and T. mutans which can both be found in over 60% of the calf population by 51 weeks. While others grow more gradually, A.phagocytophilum remains at 0% throughout the observation period.*

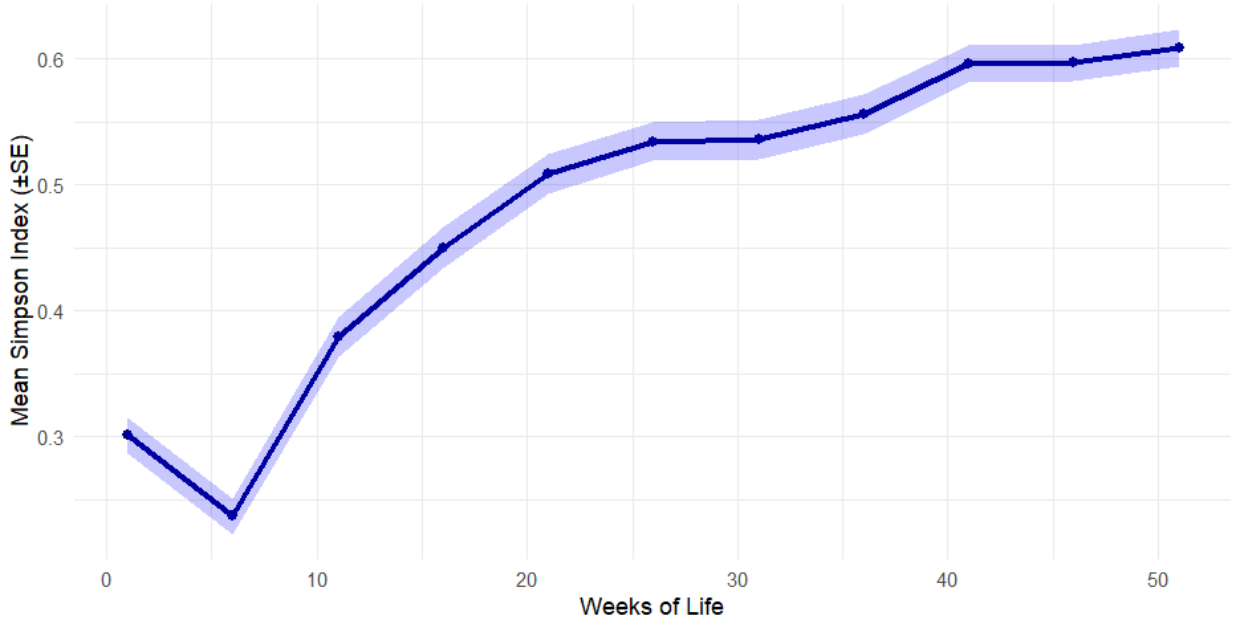


Figure 7: Simpson diversity index of haemopathogen infections across calf age (weeks of life). *The line shows the mean Simpson index per week, with shaded areas representing standard error. Higher values indicate greater infection diversity within the calf population.*

3.1.3 Pathogen load and risk profiles

In order to investigate risk factors associated with death due to ECF various association analyses of pathogen load were carried out, *T. parva,* all *Theileria* species, and all *Anaplasma* species included in the study all shown in Figure 8. Results based on the Wilcoxon test revealed that all were associated with ECF related deaths. *T. parva* lifetime burden was strongly associated with mortality ($p < 2.2 \times 10^{-16}$). Calves that died had a markedly higher lifetime pathogen load (median: 22,115 reads; IQR: 0–84,488) than those that survived (median: 0 reads; IQR: 0–4,577) (Figure 8A). In contrast, the combined burden of all Theileria species was not significantly associated with mortality ($p = 0.87$). Median loads were in fact higher in surviving calves (100,501 reads; IQR: 5,856–403,953) compared to deceased calves (82,420 reads; IQR: 15,886–338,458) (Figure 8B). A significant association was also observed for *Anaplasma* species ($p < 2.2 \times 10^{-16}$). Surviving calves had substantially higher pathogen loads (median: 274,687 reads; IQR: 93,310–592,446) compared to those that died (median: 25,790 reads; IQR: 0–189,450) (Figure 8C).

Further pathogen load investigation was done into the association between host, associated sublocations and calf sex and how they impacted *T. parva* load in Figure 9. No significant differences in *T. parva* load were observed between male and female calves following a Wilcoxon test ($p = 0.9834$). Median loads were 0 for both sexes, with overlapping interquartile ranges (F: 0–5,006; M: 0–5,849). In Figure 10 a more in depth study of haemopathogen to calf sex association was done using a Kaplan Meier plot. Male calves exhibited lower survival probabilities throughout the 51-week period; however, as the y-axis only displays probabilities above 0.75, the apparent difference may be visually exaggerated. A chi-squared test revealed no statistically significant differences in infection rates between sexes. Referring back to Figure

9, a comparison between sublocations and *T. parva* load showed a positive association through a Kruskal-Wallis test indicating strong spatial heterogeneity. A full breakdown of the number of calves sampled in each sublocation can be found in Appendix ATable1. Median loads ranged from 0 in most sublocations to 19,384 in Magombe East. Building on the significance of sublocation, Figure 11, which illustrates mortality percentages by sublocation, shows a pattern consistent with these findings. While mortality percentages varied significantly across sublocations, Magombe East had the highest number of observed infections relative to expectation (6 vs. 1.15 expected). Elevated infection-related deaths were also observed in Bujwanga and Bumala A, while sublocations such as Bukati, Igero, Karisa, and Simur East experienced fewer deaths than expected

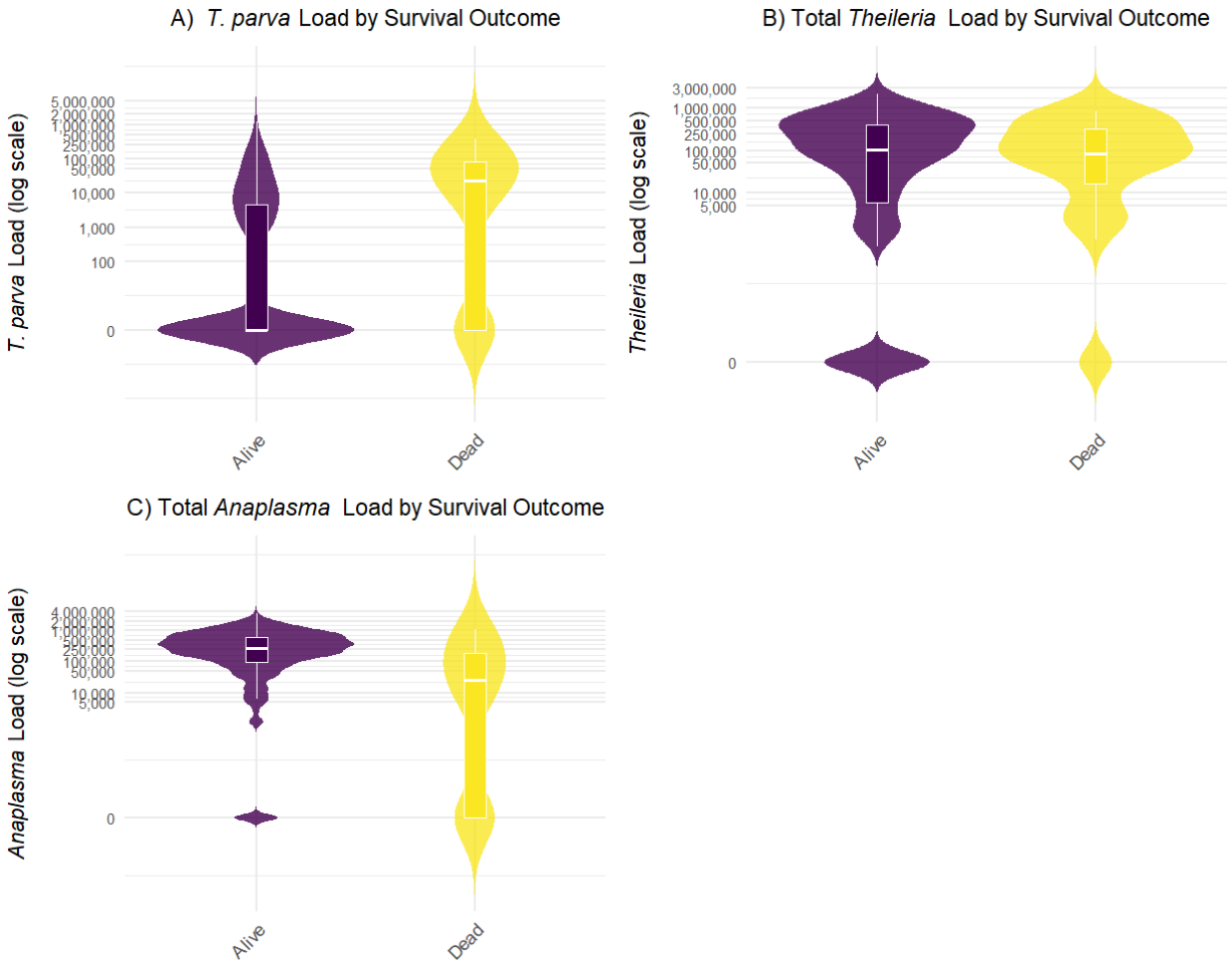


Figure 8: Violin plots comparing lifetime pathogen load seen in calves that lived the full 51 weeks of study and those that died of ECF during the study for (A) T. parva (B) Theileria and (C) Anaplasma. *(A) T. parva load compared to survival outcome.*

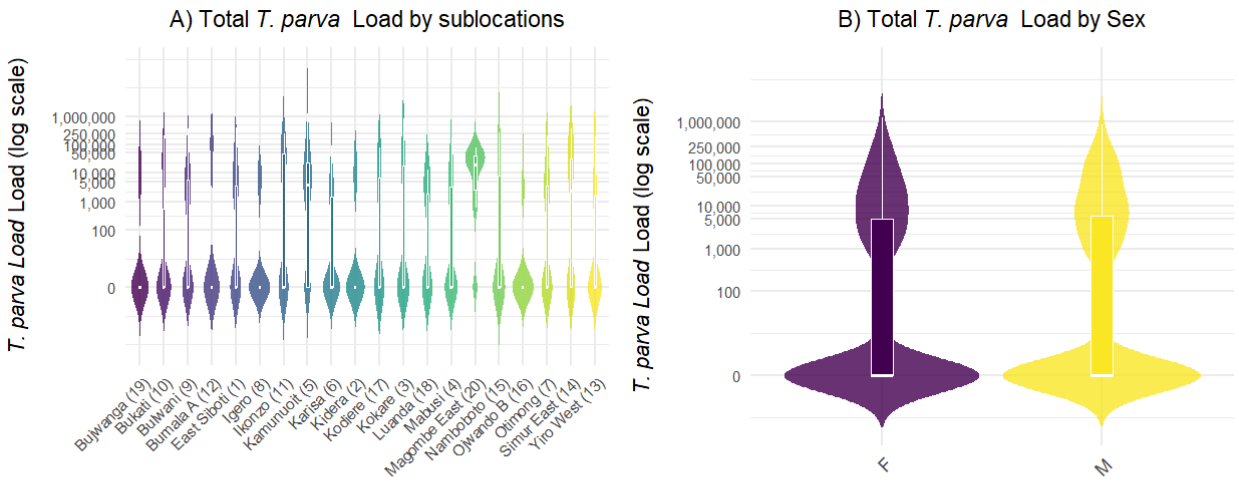


Figure 9: Violin plots showing the distribution of T. parva lifetime load across key biological and ecological variables. *(A) T. parva load compared to 20 chosen sublocations in Kenya. Variation in concentrations of read count samples can be seen throughout the sublocations with particular locations favouring larger concentrations of read count samples around the 50,000 mark such as Magombe East and Kamunuoit. Whereas those such as Igero see a much larger concentration of read counts around the 0 mark for T. parva load(B) T. parva load compared to calf sexes. This violin plot shows largely similar plots between the two sexes with similar median marks and identical areas of higher read count concentration samples.*

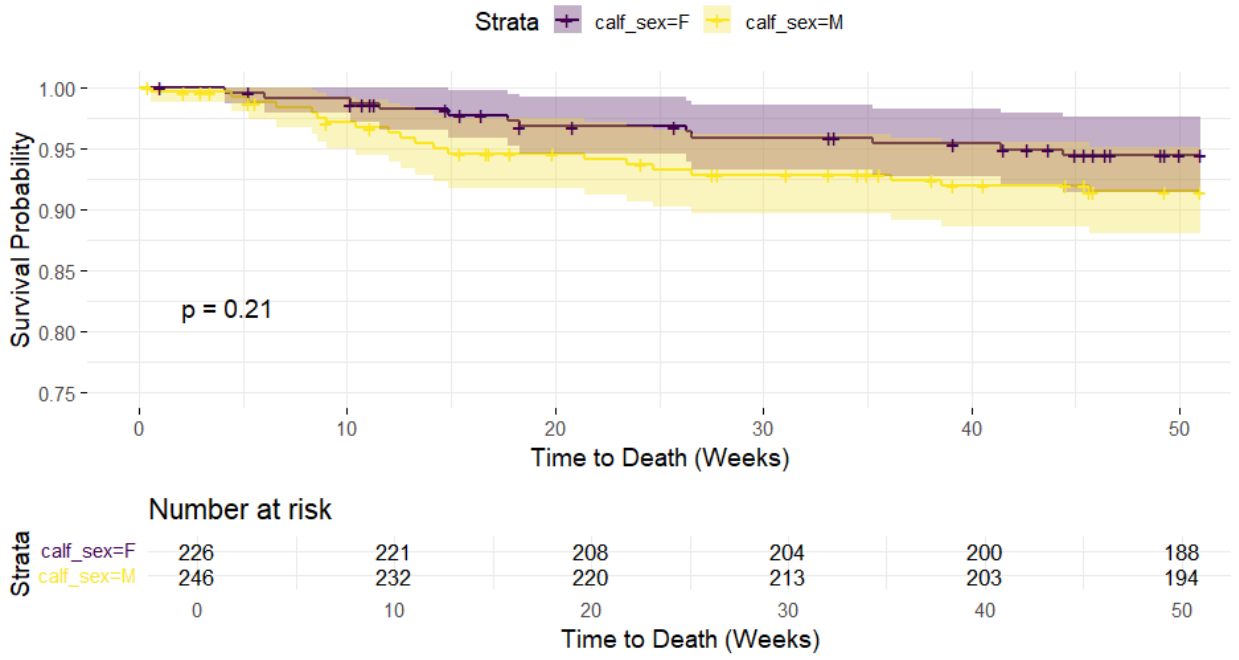


Figure 10: Kaplan-Meier survival curve showing the impact of calf sex on mortality from East Coast fever over a 50-week follow-up period. *(log-rank test, p = 0.21). The number of calves at risk is indicated below the plot at 10-week intervals.*

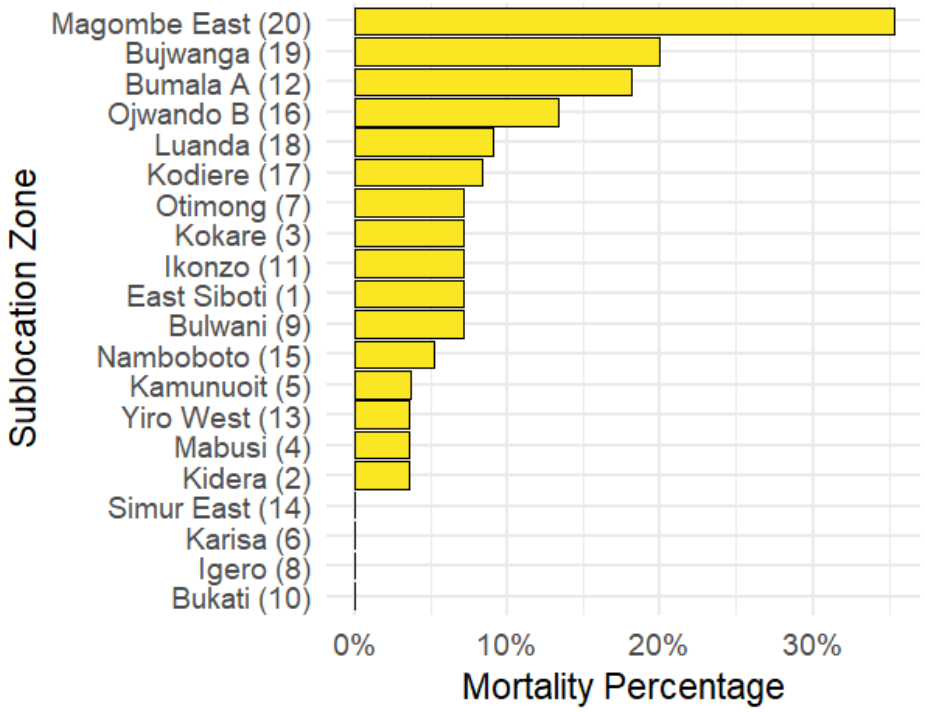


Figure 11: Mortality percentage of calves across sublocations. *Bars represent the percentage of deaths among cattle in each sublocation. Refer toTable1 of appendix A for sample sizes from each sublocation.*

A comparison of the effect of infection time on disease outcome was investigated in Figure 12 where a Kaplan Meier survival analysis was used to estimate risk difference if calves were infected with *T. parva* before or after the 25 week mark. A significant association was found between early infection exposure and a lower survival probability (Log-rank test, p = 0.033). This analysed the difference between these two haemopathogen groups seen in total sequencing reads, reads with primer sequences, and filtered reads per sample. Across all measures the *Anaplasma/Ehrlichia* primer set generally yielded higher read counts than the *Theileria/Babesia* set.

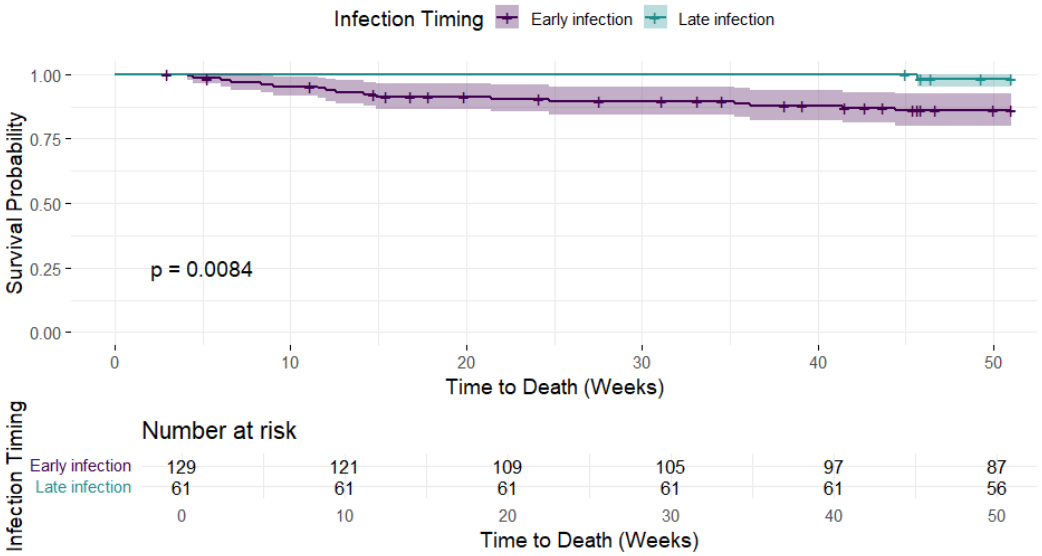


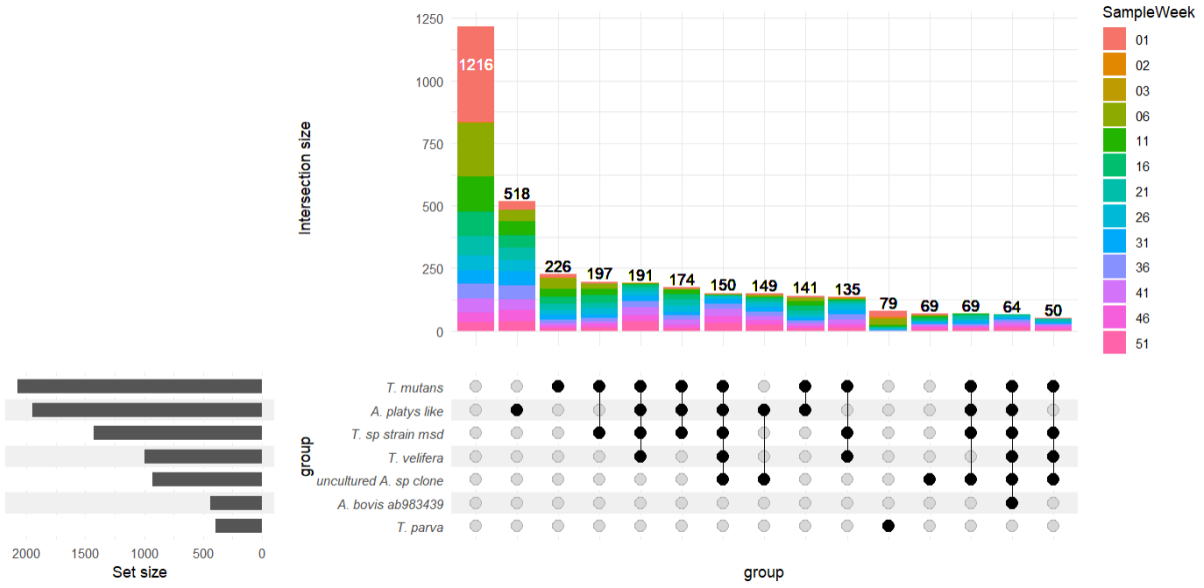
Figure 12: Kaplan Meier survival curves comparing calves infected with T. parva early (≤25 weeks of age) versus late (>25 weeks of age).

3.2 Infection dynamics (Co-infections & order effects)

To investigate co-infection dynamics further, Table 1 within appendix B shows the proportion of calves within the dataset with no infections, a single infection, dual or three or more infections at any point in the study, the majority of the cohort (84.3%) harboring 3 or more infections. Evidence that high co-infection count correlates with high overall pathogen lifetime burden can be seen in Figure 1 of appendix B where a strong positive correlation was found using a Spearman's test (CI: 0.6454, 0.7443) between the number of co-infecting pathogens and the total pathogen load per calf. Table 2 of appendix displays the frequency and mortality rates associated with unique haemopathogenic co-infection combinations. The highest mortality rate in this analysis is associated with a solitary *T. parva* infection, which sees 83.3% of those calves die, second to which are calves infected with no pathogens at all.

Figure 13 highlights the most frequent co-infection patterns among haemopathogen species detected in over 50 calves. Figure 13-A shows this divided further by sample week. It can be seen that a small subset of combinations accounted for a large proportion of calves, indicating that while most animals experienced highly individualised infection trajectories, certain co-infection profiles were recurrent and potentially epidemiologically significant. Notably, *T. mutans* and *A. platys-like* were the most frequent components of multi-pathogen infections, often appearing alongside *T. velifera* and *T. parva.* When stratified by survival outcome in 13 B, certain co-infection profiles were disproportionately associated with mortality. For example, *T. parva* mono-infections were strongly linked to death, while some multi-pathogen profiles, particularly those including *T. mutans* and *T. velifera,* appeared more common in surviving calves. Another 2 upset plots can be found in Figure 2 and 3 of Appendix B, the first of which shows co-infection patterns among *Theileria* species and the second of which concentrates on *Anaplasma* species. These graphs follow the expected pattern of signs of higher risk of mortality in co-infections including *T. parva*.

A



B

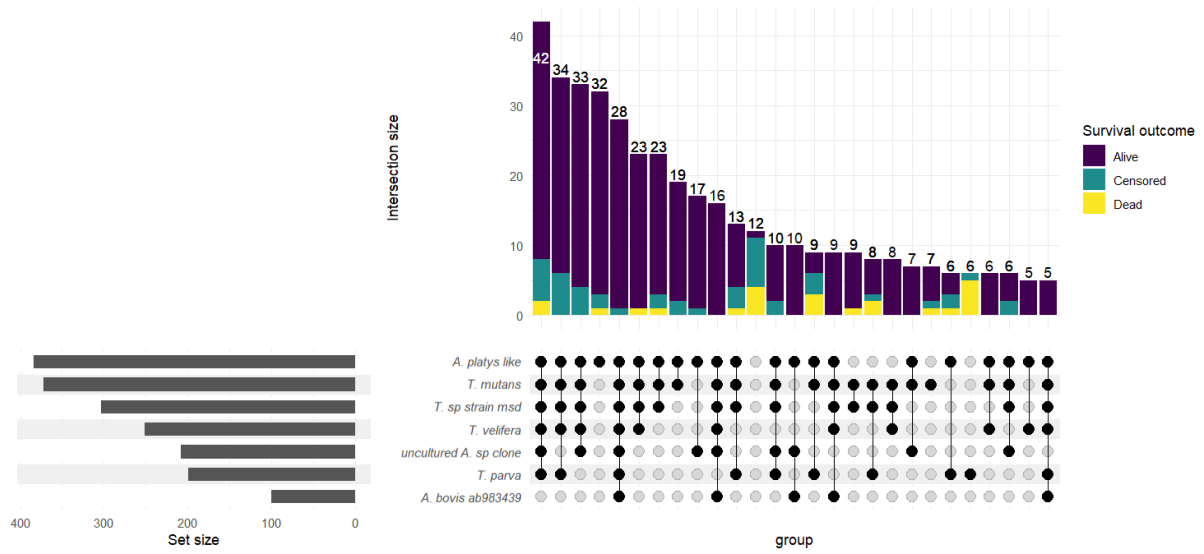


Figure 13: UpSet plot restricted to haemopathogen species detected in more than 50 calves across the cohort.

An investigation into the association seen between *T. parva* and mortality found mortality rate (Appendix BTable3) was dramatically higher in those with *T. parva*-only infection (83.3%) compared to those infected with *T. parva* and other co-infections (8.8%). The *T. parva* group had markedly greater odds of mortality than those without *T. parva* (OR= 121.5835 (95% CI 12.2–5942.3)). This was also the case between calves with *T. parva* only infections that showed a greater chance of mortality versus *T. parva* + co-infection (OR= 49.45(95% CI 5.1–2420.7)). While statistically significant, the magnitude of risk is uncertain due to the wide confidence intervals seen in the *T. parva* only group.

An important co-infection pattern investigated in this study was the relationship between *T. parva, T. velifera* and T. *mutans*. The effect of exposure order on disease outcome was analysed using temporal pathogen detection data and survival outcomes. Calves were grouped based on which pathogen was detected earliest: '*T. parva* first', '*T. mutans/ T. velifera* first', or 'no exposure to all 3 species' throughout the 51 weeks of study. A comparison between individual *Theileria* species and the probability of remaining uninfected by the three *Theileria* species at the 51 weeks mark can be seen in Figure 4 of appendix B. Calves were least likely to remain uninfected by *T. mutans* as the study went on, while *T. parva* and *T. velifera* exhibited more gradual infection acquisition curves. A significant difference was found between the 3 species and remaining uninfected ($p < 0.0001$).

When comparing *T. parva* load between survivors and non-survivors in Figure 14, there was no significant difference in the "T. *Parva* First" group, but a significant difference in the "T. *mutans/T. velifera* First" group, with non-survivors exhibiting higher median *T. parva* burdens. A two-way ANOVA showed a strong main effect of exposure order on *T. parva* burden, with a smaller significant association seen between exposure order and disease outcome. Kaplan–Meier survival analysis done in Figure 15 revealed a significantly lower survival probability in calves infected with *T. parva* before any other *Theileria* species (log-rank test, p= <0.0001). Contrastingly, calves who were earlier infected with *T. mutans* or *T. velifera* had notably higher survival rates, comparable to those who experienced no *Theileria* infection.

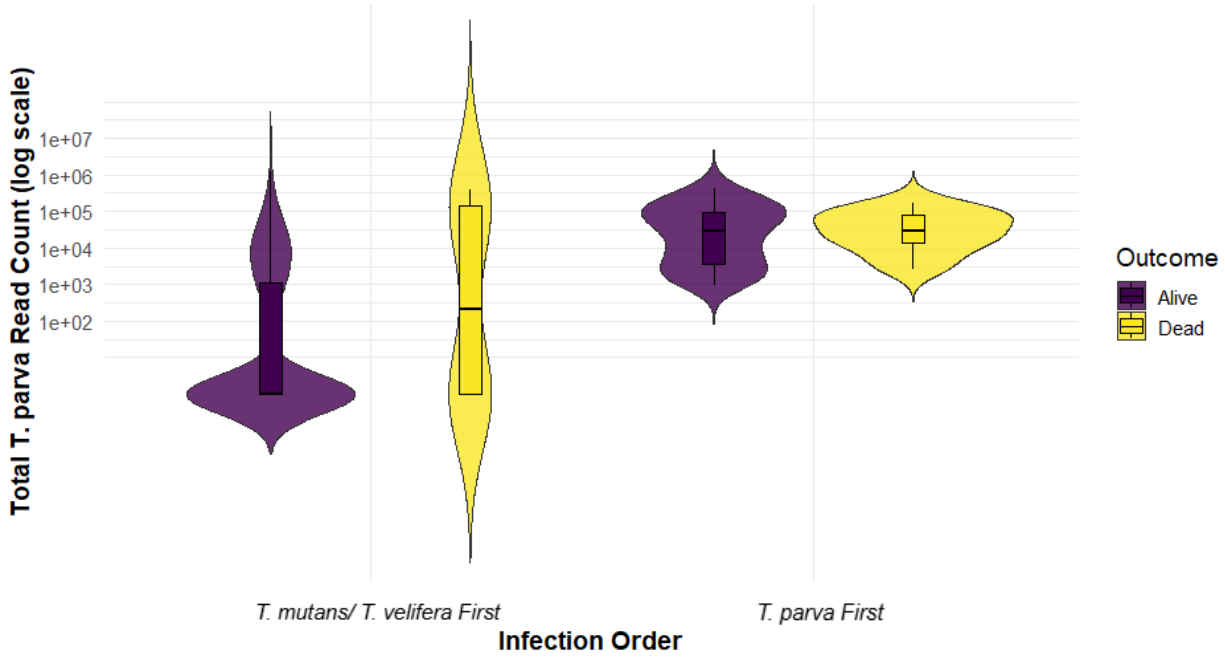


Figure 14: A comparison of T. parva read count associated with surviving (red) and non-surviving (blue) calves in the T. parva group (right) and in the calves that experienced a T. mutans and/or T. velifera infection before a T. parva infection (left). *In the mutans velifera first group the calves that died are made up of much less concentrated read count samples, this and the longer tails extending upwards and downwards provides evidence for a much larger wider spread of variance in read counts. The median read count for the dead calves is also much higher than that of those who lived. When looking at the parva first group, the concentration of read count samples and variance of read counts is much smaller with largely similar median counts.*

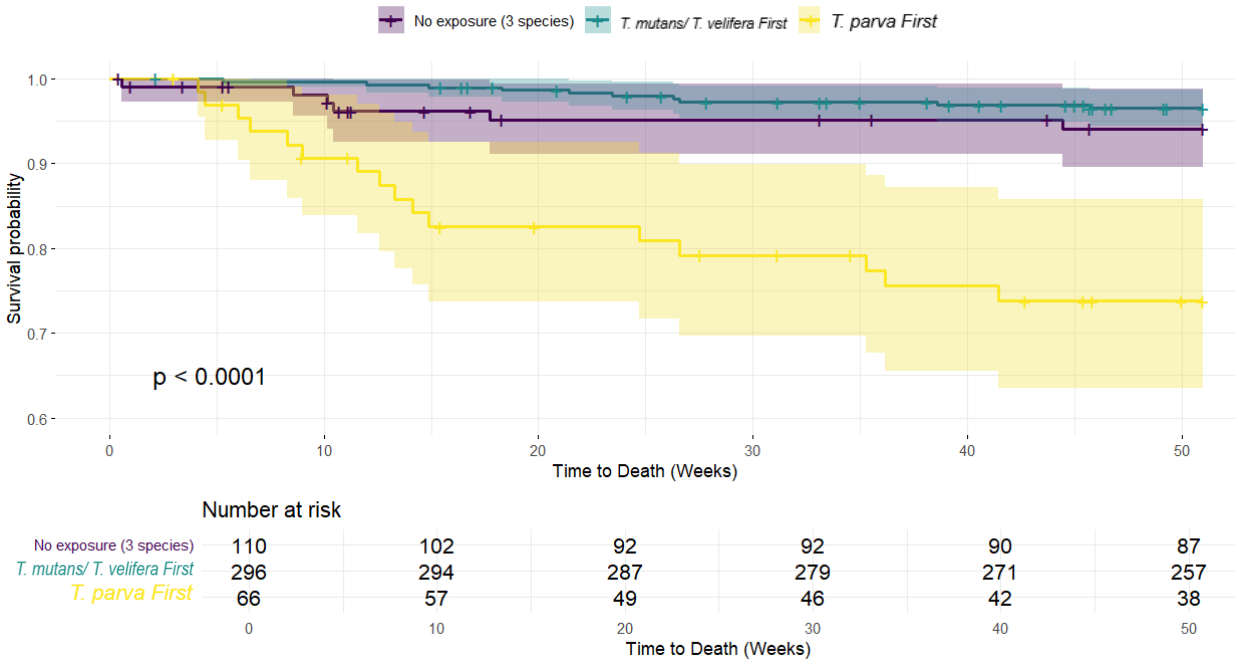


Figure 15: Kaplan Meier survival curve showing the impact of exposure order on mortality from East Coast fever over a 51-week follow-up period. *The number of calves at risk is indicated below the plot at 10-week intervals. Y axis beginning at a survival probability of 0.6.*

3.3 Assessment of the role of inherited genetic tolerance on the risk of mortality.

To understand the distribution of different genotypes across the calf dataset, Table 1 of Appendix C provides a comprehensive breakdown of calf numbers with each genotype and disease outcome. Furthermore,Table 2 in appendix C shows a reduced calf population of the 472 calves for which we have haemobiome data linked to their genotype. This data showed *T. parva* detection in 44.4% of CC calves, 40.1% of CT calves, and 39.6% of TT calves. These comparable exposure levels across genotypes indicate that TT calves were not simply unexposed and all calves experienced similar levels of environmental challenge. An investigation into the association between pathogen load and genotype in Figure 1 of Appendix C, although revealing no statistically significant difference between genotype groups following a Kruskal–Wallis test (p-value = 0.1978), provided visual evidence for TT calves commonly exhibiting lower median pathogen loads than the calves with a CC genotype.

A Kaplan–Meier survival analysis shown in Figure 16 showed that survival probability remained highest in TT calves, while CC and CT calves followed more pronounced declines over the 51-week follow-up period. However, this difference was not statistically significant. In order to address the quasi-complete separation caused by the lack of deaths in the TT group a Firth's penalized Cox regression was used in table 2, the model revealed a significantly reduced hazard of death in TT calves compared to CC (*HR = 0.14, 95% CI: 0.001–0.99)*, but no significant difference in CT calves.The overall model did not reach statistical significance (LRT p

= 0.12). Penalized likelihood estimation was used to address potential separation or small sample bias.

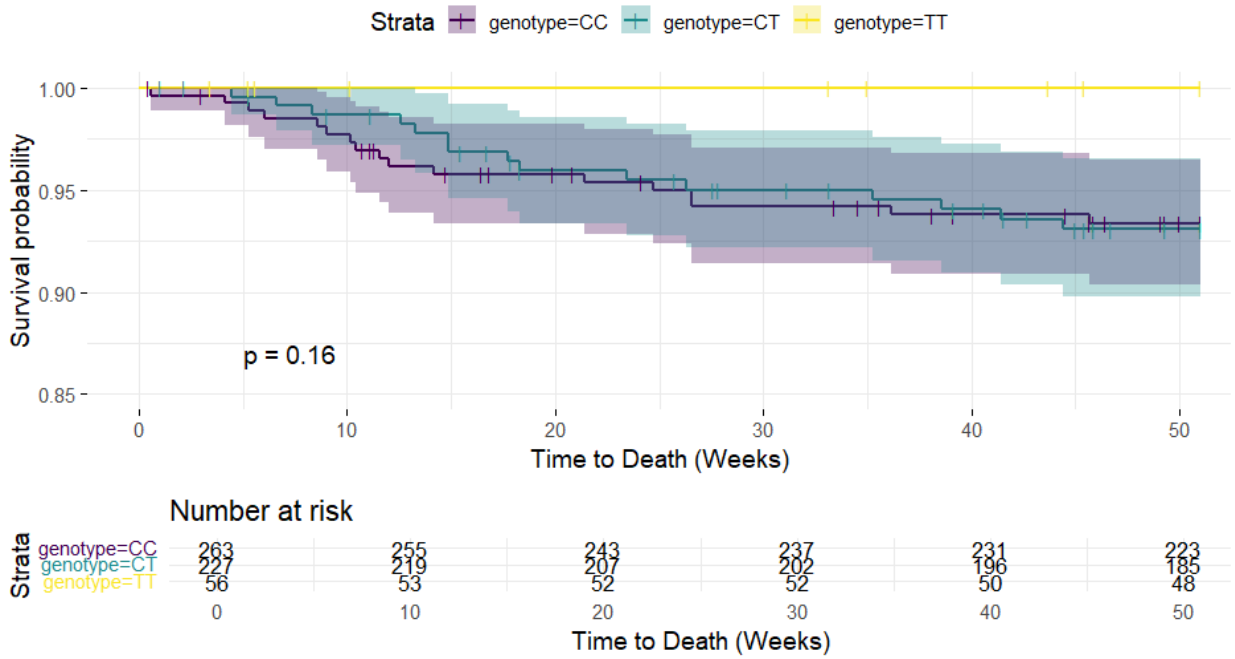


Figure 16: A Kaplan Meier survival analysis showing the impact of FAF1B genotype on mortality from East Coast fever over a 51-week follow-up period. *Survival probability remained high for those calves with a genotype of TT while steeper and similar falls in survival probability were seen in those with CC and CT genotypes. No clear statistically significant difference between genotype and survival outcome was found (log-rank test, p = 0.16). The number of calves at risk is indicated below the plot at 10-week intervals. Y axis beginning at a survival probability of 0.85.*

Table 2: Firth Penalized Cox Regression Results used to assess the effect of genotype on survival, correcting for quasi-complete separation in the TT group.

Genotype	Hazard ratio	95% CI	p-value
CT vs CC	1.03	0.51 – 2.04	0.932
TT vs CC	0.14	0.001 – 0.99	0.048

3.4 Identifying key predictors to inform interventions

The final Cox-Firth penalised regression model in Figure 17 was carried out with the aim to identify several significant predictors of calf mortality. As shown in Figure 17, a greater *T. parva* load was strongly associated with an increased hazard of death, Contrastingly a higher *Anaplasma* load was significantly associated with a lower mortality risk. Exposure order was further shown to play a major role with calves infected with *T. parva* first having a substantially

higher hazard of death compared with those where *T. mutans* and/or *T. velifera* preceded *T. parva*. This Cox-Firth model is also indicative of higher co-infection counts being statistically associated with lower mortality hazards.

Host genotype showed significant effects in the univariate analysis and although genotypes TT and CT conferred a decreased mortality risk, TT more so. The large confidence intervals in the case of TT genotype indicate a high level of uncertainty after controlling for the other variables.

The AIC model comparison (Table 3) showed the full model (AIC = 267.21) provided the best overall fit, while exclusion of exposure order (ΔAIC = 16.97), co-infection count (ΔAIC = 8.6), or genotype (ΔAIC = 3.97) all substantially reduced model performance. Removing specific pathogens also worsened model fit, particularly for *Anaplasma* (ΔAIC = 16.44) and *T. parva* (ΔAIC = 5.88). In contrast, removing all *Theileria* species except *T. parva* did not meaningfully degrade the fit (ΔAIC ≈ 0.02).

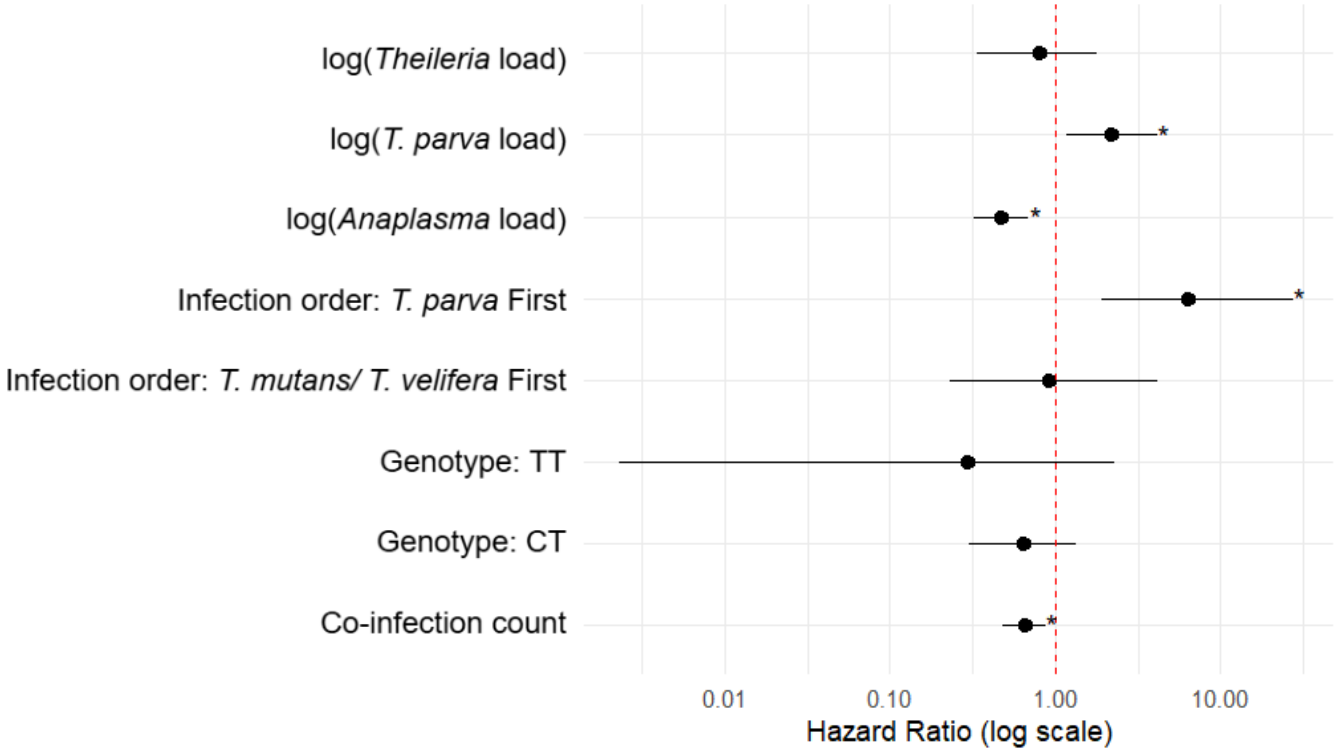


Figure 17: Plot of hazard ratios (HR) from the final Cox–Firth penalised regression model assessing predictors of calf mortality (n = 472, events = 32). *Black filled points represent HR estimates, black horizontal bars indicate 95% confidence intervals (CI), and the dashed vertical line represents the null value (HR = 1). HR > 1 indicates an increased hazard of death while HR < 1 indicates a reduced hazard. Significance is indicated by asterisks (*p < 0.05, **p < 0.01, ***p < 0.001).*

Table 3: Comparison of Cox-Firth penalised regression models for predictors of East Coast fever outcomes. Models compared using Akaike Information Criterion (AIC), and difference in AIC relative to the full model (△AIC).

Model	AIC	△ AIC
Full model	267.21	0
No genotype	271.19	3.97
No exposure order	284.18	16.97
No co-infection count	275.81	8.60
No *T. parva*	273.1	5.88
No *Anaplasma*	283.65	16.44
No *Theileria*	267.23	0.02

4. Discussion

4.1 Descriptive analysis

4.1.1 Cohort survival outcomes

The histogram in figure 4 shows that the majority of deaths as a result of ECF happen withing the first few months of life, this aligns with heightened *T. parva* susceptibility in calves at a young age. Reflecting expected immunological immaturity and early exposure to infected ticks. However, the majority of calves make it to the 51 week mark, displaying evidence of effective immune responses or reduced exposure to some areas of the calf cohort.

4.1.2 Pathogen prevalence over time

Figure 5 and 6 highlight the substantial burden of haemopathogen infections in East African smallholder cattle systems, with nearly all calves experiencing at least one infection during their first year of life. When examining haemopathogen prevalence over time, it was observed that an increase in *Theileria* and *Anaplasma* levels around 10 weeks coincided with higher ECF-associated mortality. This pattern suggests that rising *T. parva* burdens in the population may contribute to elevated mortality risk. While several haemopathogens were detected at high prevalences and other causes of death accounted for in Appendix A Figure 1 such as haemonchosis and heartwater, the fact that 32.7% of calf deaths were attributable to ECF highlights *T. parva* as a key pathogen of concern in this setting. This is consistent with existing literature, which identifies *T. parva* as the most pathogenic of the tick-borne haemopathogens in East Africa (Bishop et al., 2020; Morrison et al., 2020). However, the results indicate that *T. parva* prevalence measured as the percentage of calves PCR-positive at a given time did not

show a marked increase at 10 weeks, contrasting with the patterns observed for other haemopathogens. However, this prevalence method may not fully capture the true pathogen dynamics if calves experience transient infections and then test negative at later samplings.

The increase in Simpson diversity with calf age indicates a shift from dominance by fewer haemopathogens, to a more broadly distributed group of pathogens. This is presumably due to an increased exposure to a wider range of pathogens as the calf ages through continued exposures to novel ticks and other vectors over time, resulting in a more complex co-infection profile in the older animals. The high prevalence and complexity of co-infection patterns observed in this study underscore the multifactorial nature of haemopathogen disease dynamics in endemic regions. These findings align with previous work suggesting that polyparasitism is not a rare occurrence but rather the standard in smallholder systems (Woolhouse et al., 2015).

The findings also highlight the heavy and complex infection pressure that calves are facing in their first year and set the stage for examining which factors explain survival differences. Not all infected calves succumbed suggesting additional host, environment and pathogen-related factors shape the immortality risk, this is explored in the following section.

4.1.3 Pathogen load and risk profiles

Having established *T. parva* as a pathogen of major concern, the next step is to understand how additional host, environmental, and pathogen-related factors shape survival outcomes. Pathogen load emerged as a key determinant: calves that died of ECF carried substantially higher *T. parva* lifetime burdens compared with survivors, reinforcing the strong link between parasite load and mortality risk. By contrast, surviving calves tended to harbour higher *Anaplasma* loads, suggesting a potential protective effect. This raises the possibility that some haemopathogens may mitigate rather than exacerbate disease outcomes, either through immune modulation or competitive interactions.

The lower death rates than expected in several of the sublocations suggests the influence of localised protective factors, whether environmental, immunological, or management-related; it means disease pressure is unevenly distributed in this landscape. Such heterogeneity has been previously documented in the literature (Deem et al., 1993) with thorough reports of the impact factors such as climate, vegetation, and livestock practices can have on tick diseases such as ECF. Notably, Magombe East experienced over five times the expected number of deaths, pointing to a convergence of high infection pressure and poor survival outcomes, likely due to intense *T. parva* transmission, limited veterinary support, or low baseline immunity; however the specific reasons are not yet clear. The stark differences in disease pressures we can see between sublocations highlights the need for geographically tailored control strategies that prioritises surveillance and specific resource allocation such as focal vector control to the high-risk sublocations.

The timing of infection was found to be a critical aspect of *T. parva* related mortality. While prior studies have reported mixed age-dependent mortality patterns (Koch et al., 1990; Nicholson et al., 2019), this analysis, restricted to calves under one year, indicates that mortality is greatest

when *T. parva* infection occurs before 25 weeks of age. This is biologically plausible, given their less mature immune systems and potentially diminished capacity to mount effective responses particularly in the absence of maternal antibody protection or prior exposure. The increase in *T. parva* death when infection happens under 25 weeks highlights a critical intervention window. Calves under the age of 6 months therefore should be prioritized with treatment such as vaccination, strategic acaricide use, or passive immunity support as well as monitoring and surveillance.

Finally, the absence of any significant associations between sex and infection prevalence or survival outcome supports the hypothesis that exposure and susceptibility in this setting are independent of sex-based biological differences, as supported by (Nyabongo et al., 2021) which found no association between sex and prevalence of ECF.

While pathogen load, geography, and timing provide useful explanations for variation in mortality, these factors are less amenable to direct control. In contrast, if infection sequence or host genetic variation are shown to meaningfully alter outcomes, they offer pathways for targeted and sustainable interventions. For this reason, the next sections focus on exposure order and genetic tolerance as central questions of this study.

4.2 Infection dynamics (Co-infections & order effects)

The co-infection patterns observed in this study suggest that, despite the individuality of infection trajectories, a small number of recurrent profiles have clear epidemiological significance. The most prominent example is the contrast between *T. parva* mono-infections, which were associated with extremely high mortality, and multi-pathogen infections including *T. parva together with T. mutans* and/or *T. velifera*. These infections were disproportionately observed among surviving calves. Statistical testing supported this finding, showing that calves with *T. parva* only infections had much greater odds of death than those co-infected with *T. parva* and other haemopathogens. This consistent trend points to a compelling biological signal: co-infection with other haemopathogens may attenuate the lethality of *T. parva.*

Furthering the analysis of this hypothesis the significance of *T. mutans* and *T. velifera* should be understood and temporal ordering of infections in previous literature has shown great association with clinical outcome, this supports previous hypotheses from immuno-epidemiological models that infection sequence can shape immune responses and pathogen competition (Balmer & Tanner, 2011; Cobey & Lipsitch, 2013). In this analysis, calves infected with *T. parva* before other *Theileria* species experienced significantly poorer survival, whereas those first exposed to *T. mutans* or *T. velifera* had survival probabilities comparable to calves without *Theileria* infection. This supports prior studies where non-pathogenic *Theileria* species have been shown to modulate immune responses, possibly through early activation of innate or cross-reactive adaptive immunity (Nene et al., 2016; Woolhouse et al., 2015). These less virulent infections may act as immunological primers, enhancing host resistance to subsequent challenge by more pathogenic species like *T. parva.*

When considered alongside parasite load, the results provide further support for a protective effect of early exposure to less pathogenic species. Among calves in the “T. *mutans/T. velifera* first” group, non-survivors carried higher *T. parva* lifetime burdens than survivors, suggesting that although these calves eventually succumbed, early infection with *T. mutans* or *T. velifera* may have slowed disease progression and allowed for longer survival. This is reflected in the later timing of death in this group compared to calves infected with *T. parva* first (Appendix B, Figure 5). The broader range of *T. parva* loads observed in the *“T*. *mutans/T. velifera* first” non-survivors may therefore reflect cumulative exposure over a longer lifespan.

By contrast, in calves infected with *T. parva* first, parasite load did not significantly distinguish survivors from non-survivors, implying that once early exposure to *T. parva* occurs, outcome is determined more by host susceptibility than by parasite replication level. Together, these findings indicate that exposure order influences not only mortality risk but also the dynamics of pathogen replication and timing of death. This strengthens the hypothesis that early infection with less pathogenic *Theileria* species may induce a form of cross-protective immunity or tolerance. If substantiated, this would highlight infection sequence as a potentially controllable risk factor, opening up avenues for novel interventions such as vaccines, controlled exposure strategies, or targeted management practices aimed at modulating the order of pathogen encounter.

4.3 Assessment of the role of inherited genetic tolerance on the risk of mortality.

Beyond infection dynamics, host genetics emerged as a critical factor influencing survival outcomes, with analysis of the *FAF1B* locus providing insight into potential tolerance mechanisms against *T. parva*. These results provide moderate but compelling evidence that the *FAF1B* TT genotype confers protection against ECF-related mortality. Although Kaplan–Meier and log-rank tests did not reveal statistically significant group-wide survival differences and confidence intervals remained wide due to the small TT sample size, the absence of deaths among TT calves and the significantly reduced hazard identified through Firth’s penalized Cox regression are biologically meaningful signals. Importantly, through Table 2 and Figure 1 in Appendix C it can be seen that the number of calves infected with *T. parva* and the exposure levels to *T. parva* were similar across genotypes, ruling out the possibility that TT calves simply avoided infection. This supports the interpretation that survival benefits arise from post-infection processes rather than differential exposure. Furthermore, although not a significant difference between them, the CC genotype exhibited a visibly lower survival probability than the CT genotype at several time points in the Kaplan Meier analysis, the trend of increasing protection with T allele dosage is consistent with recessive or dosage-dependent genetic resistance. These findings align with previous work implicating *FAF1B* allelic variation in modulating immune responses to *T. parva* in East African cattle (Wragg et al., 2022).

The distinction between genetic resistance and genetic tolerance is critical for interpreting these results (Howick & Lazzaro, 2017). Resistance mechanisms typically act by reducing pathogen

lifetime burden, whereas tolerance mechanisms allow infection but mitigate its pathological effects. In this study, no statistically significant differences in *T. parva* load were observed across genotypes, but TT calves tended to show lower median lifetime burdens and the distribution of pathogen loads in TT calves was more constrained, a direction of effect that could reflect improved control of infection dynamics. This pattern suggests that the TT genotype may act through tolerance mechanisms, permitting infection but reducing risk of mortality. These trends are similar to those observed in other studies of tick-borne disease tolerance in which genetic adaptations did not prevent infection but instead limited the disease severity (Howick & Lazzaro, 2017). This has major implications for control strategies in smallholder systems where vector control and vaccination coverage are incomplete: selecting for tolerance alleles such as *FAF1B* TT could help preserve host fitness under heavy infection pressure without requiring full pathogen clearance. Although larger studies would be needed to confirm statistical significance, the trends seen in this study provide early empirical support for the practical utility of genomic selection for tolerance traits such as those conferred by the *FAF1B* TT genotype.

4.4 Identifying key predictors to inform interventions

Finally, a penalized Cox-Firth regression was used to highlight the multifactorial determinants of calf mortality, integrating pathogen, infection sequence, co-infection, and host genetic effects. As expected, higher *T. parva* loads were strongly associated with increased hazard of death, whereas higher *Anaplasma* loads were linked to reduced mortality, consistent with a potential protective or tolerance-inducing role for this pathogen. Exposure order emerged as a key determinant: calves infected with *T. parva* prior to *T. mutans* or *T. velifera* faced substantially higher mortality risk, reinforcing the earlier observations that early exposure to less pathogenic *Theileria* species can mitigate disease severity. Additionally, a greater number of co-infections was associated with lower hazard, suggesting that diverse pathogen exposure may alter immune responses in ways that reduce the lethality of *T. parva*. Host genotype further contributed to survival outcomes, with TT and CT alleles conferring reduced mortality risk, particularly TT, although the wide confidence intervals reflect uncertainty due to small sample size.

When looking at the comparison of model predictors included in Table3 of the results the full model including all covariates had the lowest AIC (267.21), indicating the best fit. Removal of exposure order had the greatest change in AIC of 16.97, co-infection count impacted AIC by 8.6 and the removal of Analplasma burden impacted AIC by 16.44. These results indicate that exposure order, co-infection burden, and co-infection context are the strongest predictors of disease outcome, whereas genotype exerts a smaller, though still notable, effect. Overall, these results synthesize the study’s main findings: mortality risk is shaped by a combination of pathogen burden, the sequence of infection, co-infection patterns, and host genetics, highlighting multiple avenues, both epidemiological and genetic, for potential intervention.

4.2 Limitations

4.2.1 Methodological constraints

A key limitation associated with a study of retrospective data is that the study design and data collection protocols are fixed meaning there is limited control to further investigate variables of interest. For example in the case of co-infections timing and prevalence the fact sampling occurs every 5 weeks means there is a possibility that transient infections were missed and therefore absolute precision into infection timings and co-infection sequences is not able to be completely determined. Furthermore certain variables such as immunological status would have been useful to study the observed protective effect of co-infections and how this is mediated by immune modulation, for this cytokine profiles or immune cell counts could have been used. Extra variables such as these would have aided in refining certain interpretations. Another aspect is the definition of ‘early’ and ‘late’ infections particularly in Figure 12 was selected based on study length time and not on biological threshold, this runs the risk of oversimplifying the nature of immune maturation.

The IDEAL dataset looks at smallholder farm systems in Western Kenya specifically, this creates findings and hypotheses that may not be directly transferable to other ecological or production systems. A change in region will bring about different vector densities, breed susceptibility, management practices and access to veterinary care. In particular breed susceptibility will be relevant to the study of genetic tolerance as the dominant breed in the area of Kenya in which this study is carried out is the East African Zebu, a breed showing evidence of innate tolerance to ECF infections that runs the possibility of not being shared by crossbred or other exotic cattle breeds.

4.2.2 Data quality and analytics

The data itself contained substantial limitations, survival data was right-censored in this analysis 23 calves were not found to have a definitive death diagnosis and instead labelled as ’Unknown’, these calves were handled through censoring which potentially introduced bias in mortality estimates. In the case of genetic analysis, quasi-complete separation caused by the lack of deaths in the TT group introduced reduced statistical power. Particular statistical tests and analyses such as Kaplan Meier survival curves and Wilcoxon tests assume independence between groups and non-informative censoring. However, slight violations to these rules could be possible in this dataset due to shared environments and clustered exposures. In future studies the use of multilevel modeling or frailty terms may fit this model better. Furthermore, in the final model of the study, a Cox-Firth analysis, the inclusion of sublocations was not possible as doing so would have consumed many degrees of freedom, resulting in very wide confidence intervals because cases were so thinly distributed across sublocations. The study however is more so focussed on the predictive power of the other factors included in the study and so the omission does not impact the primary conclusions.

Another limitation relates to the link between read counts and pathogen presence. While the read counts provided by the novel haemobiome tool provide valuable insights into pathogen presence, the direct correlation they have with infection burden can be influenced by numerous factors. Variations in sequencing depth, read length, and dataset size can affect the number of reads obtained, potentially leading to discrepancies in quantifying pathogen loads and interpretation using this method should be cautious.

Future studies can address these limitations by using smaller sampling intervals and further looking into immunological and clinical assessments. Increasing sample size particularly in the case of genotypes or death events will improve statistical power and therefore accurate inferences. The integration of vector data, environmental monitoring, and host immune profiling would allow for more mechanistic understanding of outcomes.

4.3 Recommendations for future research

In terms of data collection and quality, This study relied on the data collected using the novel haemobiome tool, PCR data is limited by its ability to depend on pathogen DNA being present at the time of sampling, meaning transient or cleared infections run the risk of being missed. Future studies may choose to incorporate serology alongside PCR to capture both active and past infections, this will provide a more complex picture of infection. Beyond these methods, there is growing interest in field-deployable tools that could transform livestock disease surveillance. For instance, Loop-Mediated Isothermal Amplification (LAMP) offers rapid and cheaper detection of specific pathogens under farm conditions, while nanopore sequencing provides portable, real-time genetic information used to uncover complex co-infections and even novel pathogens. Alongside emerging CRISPR-based assays which offer more precise gene-editing systems for diagnostics, these approaches promise to improve early detection and support more effective monitoring of co-infections (Lou et al., 2022; Mukherjee et al., 2025; Winkworth et al., 2020; Zheng et al., 2023).

Building on the findings of this study, various of these hypotheses warrant further investigation. The protective effect of early *T. mutans* and *T. velifera* infection against *T. parva*-related mortality was shown to be statistically significant in this study. Whether this protection occurs as a result of mechanisms such as immune modulation such as priming or competitive exclusion could be part of a future study (Vafadar et al., 2021). Future studies may wish to investigate this further through the use of longitudinal immunological work such as tracking the presence of cytokine responses, memory T-cell activation, or antibody dynamics during co-infections.

The study suggested a strong association between host genotype in the form of a tolerance locus influencing calf survival, in order to form statistical evidence with which to infer changes in policies and protection techniques a larger sample size of calves with which to compare the pathogen profiles and clinical outcomes would be necessary. This would elucidate on the justification of marker-assisted selection to improve herd resilience to ECF, a system currently

being set up in Kenya as part of the Centre for Tropical Livestock Genetics and Health (Strategic plan 2030., 2023).

Another key finding was the relationship between environmental and demographic variables of the infections. Within these is the impact of sublocations, with certain areas of Kenya exhibiting consistent association with disease survival/mortality, these differences could be attributed to differences in vector density, microclimates or even grazing practices. This is an interesting angle which warrants further study, the use of agroecological zones instead of administrative locations and the inclusion of relevant environmental variables could provide a more ecologically meaningful understanding of these patterns. Future work could also incorporate spatial mapping and vector surveillance to strengthen conclusions. Data on tick abundance, seasonality, and acaricide resistance would aid in the surveillance of transmission hotspots. By understanding risk zones farmers can hope to control the spread of disease more effectively through geo-targeted interventions.

This study showed statistical evidence of the danger of pathogen infection before 25 weeks and how it can be detrimental to calf survival, this could be further analysed through the exploration of maternal immunity. Maternal antibodies will influence a calves susceptibility in its early life and may impact the mortality in relation to a *T. parva* infection. A longitudinal analysis looking at the presence of maternal antibodies and when they wane will provide a more detailed analysis into calf immunity and a more detailed picture of when a pathogen exposure will be most detrimental to calf survival. Interventions like timed vaccinations could be designed based on these findings.

5. Conclusion

The objective of this project was to investigate the key epidemiological factors impacting the disease outcomes of calves suffering from haemopathogens, with a particular focus on ECF, an economically devastating disease across areas of Africa. This was to be completed through the study of haemopathogen co-infection dynamics in both spatial and temporal aspects and the assessment of the role of inherited genetic tolerance by drawing on longitudinal data from the IDEAL study and other sources. The reason for this project and future implications is to inform the identification of key predictors of calf survival and inform more targeted interventions with the overall aim of reducing calf mortality and enhancing economic resilience over time.

A key finding of this research supported existing literature by demonstrating that calves are able to experience an aspect of genetic tolerance from the *FAF1B* gene, this study showed an observable reduced mortality risk in calves suffering from ECF who were homozygous for the T allele (TT) compared to those with (CT) or (CC) genotypes. These findings are important for developing current control measures where the addition of breeding strategies to select for tolerance traits could offer a low-cost potential solution.

Another insightful contribution of this study is its emphasis on infection timing, research from this study revealed that those calves infected by *T. parva* after the 25 weeks had a better chance of

survival than those infected before the 25 week mark. These findings are backed by existing knowledge of immunological and neonatal protection in early calfhood. These findings point towards the benefits of more targeted control strategies such as spatial application of treatment and vaccinations. Furthermore the application of these strategies during critical windows of a calves life would help yield maximum benefits for calf survival.

The study also confirmed the influence of prevalent co-infections with the detection of *Anaplasma, Ehrlichia,* and *Babesia* alongside *Theileria* in co-infection numbers up to triple or quadruple infections. Some of the most influential co-infections were associated with increased and decreased mortality risks, for example the addition of *Anaplasma* to an infection will provide a protective effect on a calf against ECF mortality. A key finding of this study was the protective effect associated with less pathogenic *Theileria* species such as *T. mutans* and *T. velifera* prior to *T. parva* infection. Survival analysis clearly visualized how these milder infections reduce the severity of subsequent T. parva infections, highlighting their potential role in naturally mitigating East Coast fever mortality. Accounting for multi-pathogen scenarios should be a key aspect of diagnostic and treatment protocols going forward.

In conclusion, this study provides a comprehensive investigation into the multiple layers impacting ECF related mortality in a calf cohort of 478 in Kenyan smallholder farms. By advancing the understanding of co-infections, genetics and ecological and temporal aspects to ECF, more targeted and long term sustainable interventions can be carried out. This will work to benefit the economical stability of farm owners in Kenya and wider Africa as well as broader food security and climate goals.

6. Bibliography

Abdullah, D. A., Ali, M. S., Omer, S. G., Ola-Fadunsin, S. D., Ali, F. F., & Gimba, F. I. (2019). Prevalence and climatic influence on hemoparasites of cattle and sheep in Mosul, Iraq. *Journal of Advanced Veterinary and Animal Research*, *6*(4), 492–498.

Akoolo, L., Rocha, S. C., & Parveen, N. (2022). Protozoan co-infections and parasite influence on the efficacy of vaccines against bacterial and viral pathogens. *Frontiers in Microbiology*, *13*, 1020029.

Alboukadel Kassambara, Marcin Kosinski, Przemyslaw Biecek, Scheipl Fabian. (2025, September 2). *Drawing Survival Curves using "ggplot2" [R package survminer version 0.5.1]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/survminer/index.html

Alcaraz-López, O. A., Flores-Villalva, S., Cortéz-Hernández, O., Vigueras-Meneses, G., Carrisoza-Urbina, J., Benítez-Guzmán, A., Esquivel-Solís, H., Werling, D., Salguero Bodes, F. J., Vordemeier, M., Villarreal-Ramos, B., & Gutiérrez-Pabello, J. A. (2021). Association of immune responses of Zebu and Holstein-Friesian cattle and resistance to mycobacteria in a BCG challenge model. *Transboundary and Emerging Diseases*, *68*(6), 3360–3365.

Allan, F. K., & Peters, A. R. (2021). Safety and efficacy of the East Coast fever Muguga cocktail vaccine: A systematic review. *Vaccines*, *9*(11), 1318.

Alzan, H. F., Mahmoud, M. S., & Suarez, C. E. (2024). Current vaccines, experimental immunization trials, and new perspectives to control selected vector borne blood parasites of veterinary importance. *Frontiers in Veterinary Science*, *11*, 1484787.

Animal health important for helping cut greenhouse gas emissions, new report says. (2022, July 21). Food and Agriculture Organization of the United Nations; FAO. https://www.fao.org/newsroom/detail/animal-health-important-for-helping-cut-greenhouse-gas-emissions-new-report-says/

Arvidsson, A., Fischer, K., Hansen, K., Sternberg-Lewerin, S., & Chenais, E. (2022). Diverging discourses: Animal health challenges and veterinary care in northern Uganda. *Frontiers in Veterinary Science*, *9*, 773903.

Balehegn, M., Kebreab, E., Tolera, A., Hunt, S., Erickson, P., Crane, T. A., & Adesogan, A. T. (2021). Livestock sustainability research in Africa with a focus on the environment. *Animal Frontiers*, *11*(4), 47–56.

Barbet, A. F., Yi, J., Lundgren, A., McEwen, B. R., Blouin, E. F., & Kocan, K. M. (2001). Antigenic variation of Anaplasma marginale: major surface protein 2 diversity during cyclic transmission between ticks and cattle. *Infection and Immunity*, *69*(5), 3057–3066.

Barger, I. A. (1993). Influence of sex and reproductive status on susceptibility of ruminants to nematode parasitism. *International Journal for Parasitology*, *23*(4), 463–469.

Bishop, R. P., Odongo, D., Ahmed, J., Mwamuye, M., Fry, L. M., Knowles, D. P., Nanteza, A., Lubega, G., Gwakisa, P., Clausen, P.-H., & Obara, I. (2020). A review of recent research on Theileria parva: Implications for the infection and treatment vaccination method for control of East Coast fever. *Transboundary and Emerging Diseases*, *67 Suppl 1*(S1), 56–67.

Bouchard, C., Dibernardo, A., Koffi, J., Wood, H., Leighton, P. A., & Lindsay, L. R. (2019). N Increased risk of tick-borne diseases with climate and environmental changes. *Releve Des Maladies Transmissibles Au Canada [Canada Communicable Disease Report]*, *45*(4), 83–89.

Byaruhanga, J., Tayebwa, D. S., Eneku, W., Afayoa, M., Mutebi, F., Ndyanabo, S., Kakooza, S., Okwee-Acai, J., Tweyongyere, R., Wampande, E. M., & Vudriko, P. (2017). Retrospective study on cattle and poultry diseases in Uganda. *International Journal of Veterinary Science and Medicine*, *5*(2), 168–174.

Callaby, R., Pendarovski, C., Jennings, A., Mwangi, S. T., Van Wyk, I., Mbole-Kariuki, M., Kiara, H., Toye, P. G., Kemp, S., Hanotte, O., Coetzer, J. A. W., Handel, I. G., Woolhouse, M. E. J., & De, C. (2020). *IDEAL, the Infectious Diseases of East African Livestock project open access database and biobank*.

Cattle population by country. (2025, September 6). World Population Review. https://worldpopulationreview.com/country-rankings/cattle-population-by-country

CDC. (2025, May 19). *About Anaplasmosis*. CDC. https://www.cdc.gov/anaplasmosis/about/index.html

Chase, C. C. L., Hurley, D. J., & Reber, A. J. (2008). Neonatal immune development in the calf and its impact on vaccine response. *The Veterinary Clinics of North America. Food Animal Practice*, *24*(1), 87–104.

Chepkwony, R., Castagna, C., Heitkönig, I., van Bommel, S., & van Langevelde, F. (2020). Associations between monthly rainfall and mortality in cattle due to East Coast fever, anaplasmosis and babesiosis. *Parasitology*, *147*(14), 1743–1751.

Cobey, S., & Lipsitch, M. (2013). Pathogen diversity and hidden regimes of apparent competition. *The American Naturalist*, *181*(1), 12–24.

Cook, E. A. J., Sitt, T., Poole, E. J., Ndambuki, G., Mwaura, S., Chepkwony, M. C., Latre de Late, P., Miyunga, A. A., van Aardt, R., Prettejohn, G., Wragg, D., Prendergast, J. G. D., Morrison, W. I., & Toye, P. (2021). Clinical Evaluation of Corridor Disease in Bos indicus (Boran)

Cattle Naturally Infected With Buffalo-Derived Theileria parva. *Frontiers in Veterinary Science*, *8*, 731238.

Dantas-Torres, F. (2015). Climate change, biodiversity, ticks and tick-borne diseases: The butterfly effect. *International Journal for Parasitology. Parasites and Wildlife*, *4*(3), 452–461.

De Boeck, K., Decouttere, C., Jónasson, J. O., & Vandaele, N. (2022). Vaccine supply chains in resource-limited settings: Mitigating the impact of rainy season disruptions. *European Journal of Operational Research*, *301*(1), 300–317.

de Clare Bronsvoort, B. M., Thumbi, S. M., Poole, E. J., Kiara, H., Auguet, O. T., Handel, I. G., Jennings, A., Conradie, I., Mbole-Kariuki, M. N., Toye, P. G., Hanotte, O., Coetzer, J. A. W., & Woolhouse, M. E. J. (2013). Design and descriptive epidemiology of the Infectious Diseases of East African Livestock (IDEAL) project, a longitudinal calf cohort study in western Kenya. *BMC Veterinary Research*, *9*, 171.

de Villiers, E. (n.d.). *Identification of Theileria parva vaccine candidate genes using a Bioinformatics approach*. Retrieved September 3, 2025, from https://www.lirmm.fr/france_afrique/Nairobi2007/presentations/Etienne_de_Villiers.pdf

Deem, S. L., Perry, B. D., Katende, J. M., McDermott, J. J., Mahan, S. M., Maloo, S. H., Morzaria, S. P., Musoke, A. J., & Rowlands, G. J. (1993). Variations in prevalence rates of tick-borne diseases in Zebu cattle by agroecological zone: implications for East Coast fever immunization. *Preventive Veterinary Medicine*, *16*(3), 171–187.

Down to earth: Sustainable rural transformation. (2013, August 15). IFAD. https://www.ifad.org/thefieldreport/

Dudley, W. N., Wickham, R., & Coombs, N. (2016). An introduction to survival statistics: Kaplan-Meier analysis. *Journal of the Advanced Practitioner in Oncology*, *7*(1), 91–100.

Džermeikaitė, K., Krištolaitytė, J., & Antanaitis, R. (2024). Relationship between dairy cow health and intensity of greenhouse gas emissions. *Animals: An Open Access Journal from MDPI*, *14*(6), 829.

East Coast Fever. (2015, September 1). GALVmed. https://www.galvmed.org/livestock-and-diseases/livestock-diseases/east-coast-fever/

Ebert, C. L., & Becker, S. C. (2025). Tick-borne viruses in a changing climate: The expanding threat in Africa and beyond. *Microorganisms*, *13*(7), 1509.

Ehrlichiosis. (2018, September 18). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/17958-ehrlichiosis

Ekwem, D., Morrison, T. A., Reeve, R., Enright, J., Buza, J., Shirima, G., Mwajombe, J. K., Lembo, T., & Hopcraft, J. G. C. (2021). Livestock movement informs the risk of disease spread in traditional production systems in East Africa. *Scientific Reports*, *11*(1), 16375.

Fenta, M. D., Bazezew, M., Molla, W., Kinde, M. Z., Mengistu, B. A., & Dejene, H. (2024). A systematic review and meta-analysis of contagious bovine pleuropneumonia in Ethiopian cattle. *Veterinary and Animal Science*, *26*(100410), 100410.

Firke, S. (2024, December 22). *Simple Tools for Examining and Cleaning Dirty Data [R package janitor version 2.2.1]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/janitor/index.html

Flay, K. J., Hill, F. I., & Muguiro, D. H. (2022). A Review: Haemonchus contortus Infection in Pasture-Based Sheep Production Systems, with a Focus on the Pathogenesis of Anaemia and Changes in Haematological Parameters. Animals : an open access journal from MDPI, 12(10), 1238.

Fry, L. M., Schneider, D. A., Frevert, C. W., Nelson, D. D., Morrison, W. I., & Knowles, D. P. (2016). East Coast Fever caused by Theileria parva is characterized by macrophage activation associated with vasculitis and respiratory failure. *PloS One*, *11*(5), e0156004.

Gachohi, J., Skilton, R., Hansen, F., Ngumi, P., & Kitala, P. (2012). Epidemiology of East Coast fever (Theileria parva infection) in Kenya: past, present and the future. *Parasites & Vectors*, *5*(1), 194.

Garcia, K., Weakley, M., Do, T., & Mir, S. (2022). Current and future molecular diagnostics of tick-borne diseases in cattle. *Veterinary Sciences*, *9*(5), 241.

Georg Heinze, Meinhard Ploner, Lena Jiricka, Gregor Steiner. (2023, August 11). *Cox Regression with Firth's Penalized Likelihood [R package coxphf version 1.13.4]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/coxphf/index.html

Githaka, N. W., Kanduma, E. G., Wieland, B., Darghouth, M. A., & Bishop, R. P. (2022). Acaricide resistance in livestock ticks infesting cattle in Africa: Current status and potential mitigation strategies. *Current Research in Parasitology & Vector-Borne Diseases*, *2*(100090), 100090.

Graf, J. F., Gogolewski, R., Leach-Bing, N., Sabatini, G. A., Molento, M. B., Bordin, E. L., & Arantes, G. J. (2004). Tick control: an industry point of view. *Parasitology*, *129 Suppl*, S427-42.

Hadley Wickham, Davis Vaughan, Maximilian Girlich, Kevin Ushey. (2024, January 24). *Tidy Messy Data [R package tidyr version 1.3.1]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/tidyr/index.html

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan. (2023, November 17). *A Grammar of Data Manipulation [R package dplyr version 1.1.4]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/dplyr/index.html

How Improved Livestock Health Can Reduce GHG Emissions. (2024, February 19). HealthforAnimals. https://healthforanimals.org/animal-health-in-data/sustainability/how-improved-livestock-health-can-reduce-ghg-emissions/

Howick, V. M., & Lazzaro, B. P. (2017). The genetic architecture of defence as resistance to and tolerance of bacterial infection in Drosophila melanogaster. *Molecular Ecology*, *26*(6), 1533–1546.

Irvin, A. D., & Mwamachi, D. M. (1983). Clinical and diagnostic features of East Coast fever (Theileria parva) infection of cattle. *The Veterinary Record*, *113*(9), 192–198.

J. A. W. Coetzer, R. C. T. (2005). *Infectious Diseases of Livestock, 2nd Edition*. Oxford University Press.

Kim, H.-Y. (2014). Statistical notes for clinical researchers: Nonparametric statistical methods: 2. Nonparametric methods for comparing three or more groups and repeated measures. *Restorative Dentistry & Endodontics*, *39*(4), 329–332.

Kim, H.-Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, *42*(2), 152–155.

Klein, S. L., & Flanagan, K. L. (2016). Sex differences in immune responses. *Nature Reviews. Immunology*, *16*(10), 626–638.

Klompen, H., Walker, J. B., Keirans, J. E., & Horak, I. G. (2001). The genus Rhipicephalus (Acari, Ixodidae). A guide to the brown ticks of the world. *The Journal of Parasitology*, *87*(4), 823.

Krassowski, M. (2021, December 11). *Create Complex UpSet Plots Using "ggplot2" Components [R package ComplexUpset version 1.3.3]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/ComplexUpset/index.html

Lacasta, A., Mwalimu, S., Kibwana, E., Saya, R., Awino, E., Njoroge, T., Poole, J., Ndiwa, N., Pelle, R., Nene, V., & Steinaa, L. (2018). Immune parameters to p67C antigen adjuvanted with ISA206VG correlate with protection against East Coast fever. *Vaccine*, *36*(11), 1389–1397.

Leal, B., Zamora, E., Fuentes, A., Thomas, D. B., & Dearth, R. K. (2020). Questing by tick larvae (Acari: Ixodidae): A review of the influences that affect off-host survival. *Annals of the Entomological Society of America*, *113*(6), 425–438.

Lee, S., Clémentine, C., & Kim, H. (2024). Exploring the genetic factors behind the discrepancy in resistance to bovine tuberculosis between African zebu cattle and European taurine cattle. *Scientific Reports*, *14*(1), 2370.

Lei, S., Chen, S., & Zhong, Q. (2021). Digital PCR for accurate quantification of pathogens: Principles, applications, challenges and future prospects. *International Journal of Biological Macromolecules*, *184*, 750–759.

Le-Rademacher, J. G., Therneau, T. M., & Ou, F.-S. (2022). The utility of multistate models: A flexible framework for time-to-event data. *Current Epidemiology Reports*, *9*(3), 183–189.

Lou, J., Wang, B., Li, J., Ni, P., Jin, Y., Chen, S., Xi, Y., Zhang, R., & Duan, G. (2022). The CRISPR-Cas system as a tool for diagnosing and treating infectious diseases. Molecular biology reports, 49(12), 11301–11311.

Mada, P. K., & Alam, M. U. (2025). Clostridioides difficile infection. In *StatPearls*. StatPearls Publishing.

Maharana, B. R., Tewari, A. K., Saravanan, B. C., & Sudhakar, N. R. (2016). Important hemoprotozoan diseases of livestock: Challenges in current diagnostics and therapeutics: An update. *Veterinary World*, *9*(5), 487–495.

McNeilly, T. N., & Nisbet, A. J. (2014). Immune modulation by helminth parasites of ruminants: implications for vaccine development and host immune competence. *Parasite (Paris, France)*, *21*, 51.

Meyer, D. F., Moumène, A., & Rodrigues, V. (2023). Microbe Profile: Ehrlichia ruminantium - stealthy as it goes. *Microbiology (Reading, England)*, *169*(11). https://doi.org/10.1099/mic.0.001415

Minjauw, B., & Mcleod, A. (2003). *Tick-borne diseases and poverty. The impact of ticks and tick-borne diseases on the livelihood of small*. 59–60.

Miyunga, A., Karani, B., Njeru, R., Nangekhe, G., Cook, A., de C. Bronsvoort, B. M., Wragg, D., Prendergast, J., & Toye, P. (2025). Genotyping of an Intensively Monitored Cohort of Bos indicus Cattle in Western Kenya for the FAF1B Allele Associated with Resistance to East Coast Fever (Theileria Parva Infection) with a Real-Time PCR Assay. *Time PCR Assay*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5184154

Moreau, E., & Chauvin, A. (2010). Immunity against helminths: interactions with the host and the intercurrent infections. Journal of biomedicine & biotechnology, 2010, 428593.

Morrison, W. I. (1984). Immune responses involved in immunity against. *Preventive Veterinary Medicine*, *2*(1–4), 167–177.

Mukherjee, A., Samanta, S., Das, S., Haque, M. Z., Jana, P. S., Samanta, I., Kar, I., Das, S., Nanda, P. K., Thomas, P., & Dandapat, P. (2025). Leveraging CRISPR-Cas-enhanced isothermal amplification tools for quick identification of pathogens causing livestock diseases. *Current Microbiology*, *82*(6), 260.

Murray, G. G. R., Woolhouse, M. E. J., Tapio, M., Mbole-Kariuki, M. N., Sonstegard, T. S., Thumbi, S. M., Jennings, A. E., van Wyk, I. C., Chase-Topping, M., Kiara, H., Toye, P., Coetzer, K., deC Bronsvoort, B. M., & Hanotte, O. (2013). Genetic susceptibility to infectious disease in East African Shorthorn Zebu: a genome-wide analysis of the effect of heterozygosity and exotic introgression. *BMC Evolutionary Biology*, *13*, 246.

Nanteza, A., Nsadha, Z., Nsubuga, J., Oligo, S., Kazibwe, A., Terundaja, C., Matovu, E., & Lubega, G. W. (2023). Assessment of the impact of early diagnosis and early treatment in the integrated control of East Coast fever (ECF) involving acquired immunity induced by natural infection in Ankole cattle. *Pathogens*, *12*(1), 115.

Nene, V., Kiara, H., Lacasta, A., Pelle, R., Svitek, N., & Steinaa, L. (2016). The biology of Theileria parva and control of East Coast fever - Current status and future trends. *Ticks and Tick-Borne Diseases*, *7*(4), 549–564.

Nicholson, W. L., Sonenshine, D. E., Noden, B. H., & Brown, R. N. (2019). Ticks (Ixodida). In G. R. Mullen & L. A. Durden (Eds.), *Medical and Veterinary Entomology* (pp. 603–672). Elsevier.

Nwanade, C. F., Wang, M., Li, S., Yu, Z., & Liu, J. (2022). The current strategies and underlying mechanisms in the control of the vector tick, Haemaphysalis longicornis: Implications for future integrated management. *Ticks and Tick-Borne Diseases*, *13*(2), 101905.

Nyabongo, L., Kanduma, E. G., Bishop, R. P., Machuka, E., Njeri, A., Bimenyimana, A. V., Nkundwanayo, C., Odongo, D. O., & Pelle, R. (2021). Prevalence of tick-transmitted pathogens in cattle reveals that Theileria parva, Babesia bigemina and Anaplasma marginale are endemic in Burundi. *Parasites & Vectors*, *14*(1), 6.

Obaid, M. K., Islam, N., Alouffi, A., Khan, A. Z., da Silva Vaz, I., Jr, Tanaka, T., & Ali, A. (2022). Acaricides resistance in ticks: Selection, diagnosis, mechanisms, and mitigation. *Frontiers in Cellular and Infection Microbiology*, *12*, 941831.

Onyiche, T. E., & MacLeod, E. T. (2023). Hard ticks (Acari: Ixodidae) and tick-borne diseases of sheep and goats in Africa: A review. *Ticks and Tick-Borne Diseases*, *14*(6), 102232.

Pfeiffer, M., Hoffmann, M. P., Scheiter, S., Nelson, W., Isselstein, J., Ayisi, K., Odhiambo, J. J., & Rötter, R. (2022). Modeling the effects of alternative crop–livestock management scenarios on important ecosystem services for smallholder farming from a landscape perspective. *Biogeosciences*, *19*(16), 3935–3958.

Porcelli, S., Deshuillers, P. L., Moutailler, S., & Lagrée, A.-C. (2024). Meta-analysis of tick-borne and other pathogens: Co-infection or co-detection? That is the question. *Current Research in Parasitology & Vector-Borne Diseases*, *6*(100219), 100219.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rincón, M., & Flavell, R. A. (1997). T-cell subsets: transcriptional control in the Th1/Th2 decision. *Current Biology: CB*, *7*(11), R729-32.

Sitt, T., Pelle, R., Chepkwony, M., Morrison, W. I., & Toye, P. (2018). Theileria parva antigens recognized by CD8+ T cells show varying degrees of diversity in buffalo-derived infected cell lines. *Parasitology*, *145*(11), 1430–1439.

Soliman, T., Barnes, A., & Helgesen, I. S. (2023). The hidden carbon impact of animal disease. *PloS One*, *18*(10), e0292659.

Strategic plan 2030, Centre for Tropical Livestock Genetic and Health. (2023). https://www.ctlgh.org/wp-content/uploads/2023/04/CTLGH-strategic-plan-2030-v7-final.pdf

Stuen, S. (2020). Haemoparasites-challenging and wasting infections in small ruminants: A review. *Animals: An Open Access Journal from MDPI*, *10*(11), 2179.

Surve, A. A., Hwang, J. Y., Manian, S., Onono, J. O., & Yoder, J. (2023). Economics of East Coast fever: a literature review. *Frontiers in Veterinary Science*, *10*, 1239110.

Tabor, A. E. (2022, March 9). *Anaplasmosis in Ruminants*. MSD Veterinary Manual. https://www.msdvetmanual.com/circulatory-system/blood-parasites/anaplasmosis-in-ruminants

Tamargo, J., Le Heuzey, J.-Y., & Mabo, P. (2015). Narrow therapeutic index drugs: a clinical pharmacological consideration to flecainide. *European Journal of Clinical Pharmacology*, *71*(5), 549–567.

Therneau, T. M. (2024, December 17). *Survival Analysis [R package survival version 3.8-3]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/survival/index.html

Thumbi, S. M., Bronsvoort, B. M. de C., Poole, E. J., Kiara, H., Toye, P. G., Mbole-Kariuki, M. N., Conradie, I., Jennings, A., Handel, I. G., Coetzer, J. A. W., Steyl, J. C. A., Hanotte, O., & Woolhouse, M. E. J. (2014). Parasite co-infections and their impact on survival of indigenous cattle. *PloS One*, *9*(2), e76324.

Toye, P., Kiara, H., ole-MoiYoi, O., Enahoro, D., & Rich, K. M. (2020). The management and economics of east coast fever. In *The impact of the International Livestock Research Institute* (pp. 239–273). CABI.

Understanding East Coast Fever–Tickborne Diseases in Uganda. (2025, April 3). Ticvac.Co.Ug. https://ticvac.co.ug/article/understanding-east-coast-fever%E2%80%93tickborne-diseases-in-uganda

Upton, M. (2004, February 11). *The role of livestock in economic development and poverty reduction*. https://openknowledge.fao.org/server/api/core/bitstreams/3620ee11-79f9-4e51-b5d8-1bcb82d0d4c9/content

Vafadar, S., Shahdoust, M., Kalirad, A., Zakeri, P., & Sadeghi, M. (2021). Competitive exclusion during co-infection as a strategy to prevent the spread of a virus: A computational perspective. *PloS One*, *16*(2), e0247200.

van den Brand, H. W. W. C. L. H. T. L. P. K. T. C. W. K. W. H. Y. D. D. T. (2025, April 9). *Create Elegant Data Visualisations Using the Grammar of Graphics [R package ggplot2 version 3.5.2]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/ggplot2/index.html

Vazquez-Pertejo, M. T., & Bush, L. M. (2025, January 3). *Microscopy*. MSD Manual Professional Edition; MSD Manuals. https://www.msdmanuals.com/professional/infectious-diseases/laboratory-diagnosis-of-infectious-disease/microscopy

Winkworth, R. C., Nelson, B. C. W., Bellgard, S. E., Probst, C. M., McLenachan, P. A., & Lockhart, P. J. (2020). A LAMP at the end of the tunnel: A rapid, field deployable assay for the kauri dieback pathogen, Phytophthora agathidicida. PloS one, 15(1), e0224007.

Woolhouse, M. E. J., Thumbi, S. M., Jennings, A., Chase-Topping, M., Callaby, R., Kiara, H., Oosthuizen, M. C., Mbole-Kariuki, M. N., Conradie, I., Handel, I. G., Poole, E. J., Njiiri, E., Collins, N. E., Murray, G., Tapio, M., Auguet, O. T., Weir, W., Morrison, W. I., Kruuk, L. E. B., … Toye, P. G. (2015). Co-infections determine patterns of mortality in a population exposed to parasite infection. *Science Advances*, *1*(2), e1400026.

Wragg, D., Cook, E. A. J., Latré de Laté, P., Sitt, T., Hemmink, J. D., Chepkwony, M. C., Njeru, R., Poole, E. J., Powell, J., Paxton, E. A., Callaby, R., Talenti, A., Miyunga, A. A., Ndambuki, G.,

Mwaura, S., Auty, H., Matika, O., Hassan, M., Marshall, K., … Prendergast, J. G. D. (2022). A locus conferring tolerance to Theileria infection in African cattle. *PLoS Genetics*, *18*(4), e1010099.

Wyckliff Ngetich, George Karuoya Gitau, Tequiero Abuom Okumu, Gabriel Oluga Aboge, Daniel Muasya. (2025). Morbidity, mortality, and risk factors associated with Theileria parva seropositivity in a longitudinal calf study, Narok, Kenya,. In *Veterinary and Animal Science*.

Xiao, N. (2024, June 18). *Scientific Journal and Sci-Fi Themed Color Palettes for "ggplot2" [R package ggsci version 3.2.0]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/ggsci/index.html

Yalcindag, E., Vasoya, D., Hemmink, J. D., Karani, B., Hernandez Castro, L. E., Callaby, R., Mazeri, S., Paxton, E., Connelley, T. K., Toye, P., Morrison, L. J., & Bronsvoort, B. M. de C. (2024). Development of a novel Haemabiome tool for the high-throughput analysis of haemopathogen species co-infections in African livestock. *Frontiers in Veterinary Science*, *11*, 1491828.

Zhang, Z. (2016). Semi-parametric regression model for survival data: graphical visualization with R. *Annals of Translational Medicine*, *4*(23), 461.

Zheng, P., Zhou, C., Ding, Y., Liu, B., Lu, L., Zhu, F., & Duan, S. (2023). Nanopore sequencing technology and its applications. MedComm, 4(4), e316.

Zimmer, A. J., & Simonsen, K. A. (2025). Babesiosis. In *StatPearls*. StatPearls Publishing.

Appendix A - Descriptives of the dataset

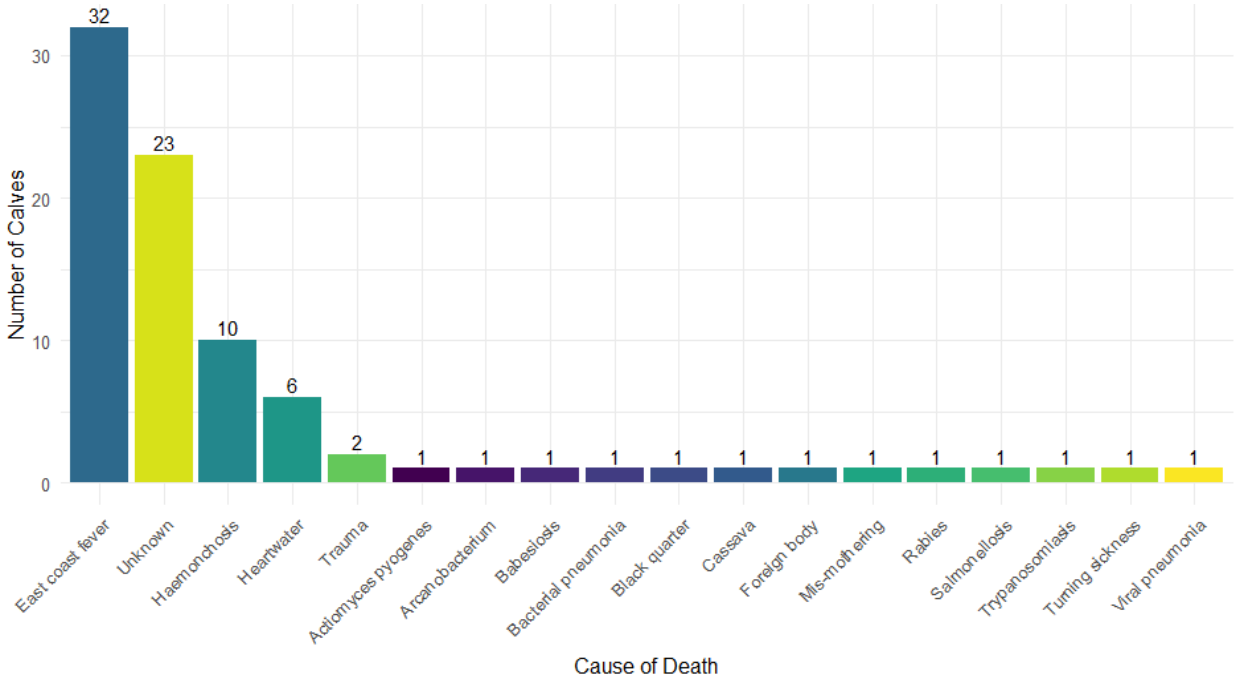


Figure 1: Number of deaths of calves as a result of specific conditions. *The largest number of deaths was due to East Coast fever (32), followed by fewer deaths due to unknown circumstances (23). All other causes of death were recorded as being 10 or less deaths per cause.*

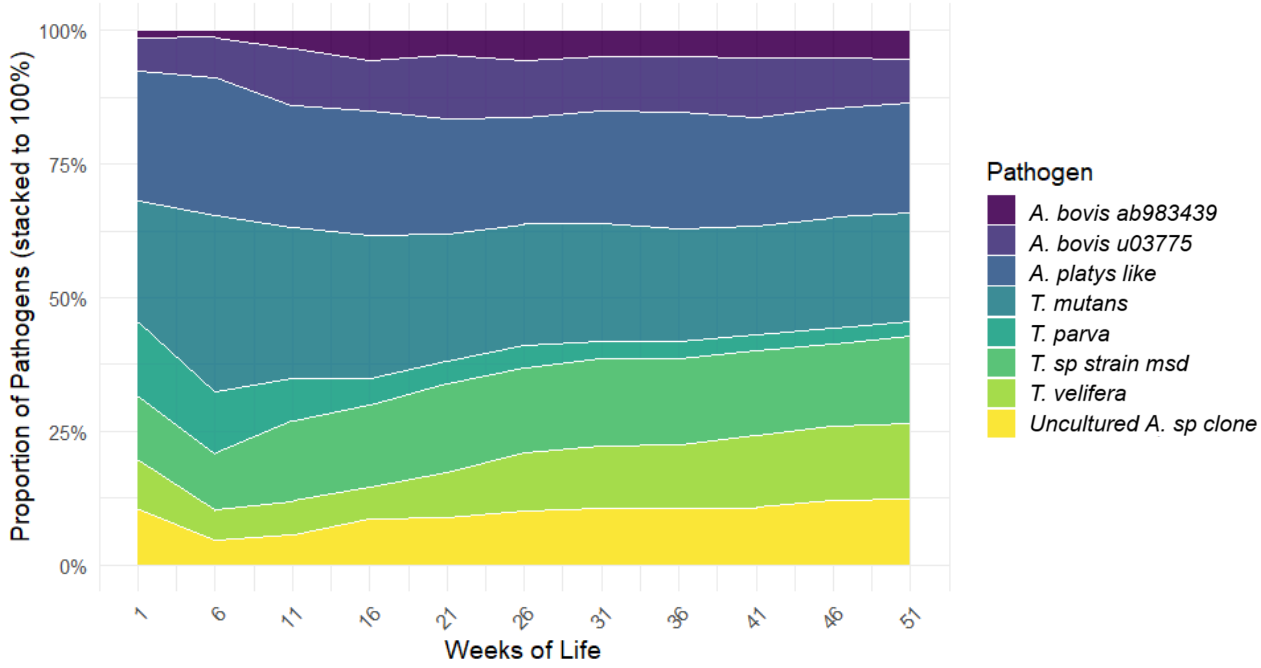


Figure 2: Proportional stacked area chart representing 100% of calves and the share of total infections that week, displayed in a ribbon graph. *Sample week shown on the x axis and proportion of calves infected in % shown on y axis. In order to not include rare pathogens this graph removes species that never exceeded 1% prevalence.*

Table 1: Number of calves sampled in each sublocation of Kenya used in this study.

Sublocation	*Number of calves*
Bukati	*28*
Bulwani	*28*
East Siboti	*28*
Igero	*28*
Ikonzo	*28*
Kidera	*28*
Kokare	*28*
Mabusi	*28*
Otimong	*28*
Yiro West	*28*
Kamunuoit	*27*
Karisa	*27*
Simur East	*27*
Bumala A	*22*
Namboboto	*19*
Magombe East	*17*
Bujwanga	*15*
Ojwando	*15*
Kodiere	*12*
Luanda	*11*

Appendix B - Co-infection prevalence and patterns

Table 1. Distribution of calves by number of haemopathogen infections detected over the study period. *ThisTable summarises the number and percentage of calves that were infected with zero, one, two, or three or more haemopathogen species at any point during the 51-week study. Larger amounts of calves are associated with more numerous co-infections.*

Infection category	Number of calves	Percentage of total
No Infections	9	1.9
Single Infections	21	4.4
Dual Infections	44	9.3
Triple or more infections	398	84.3

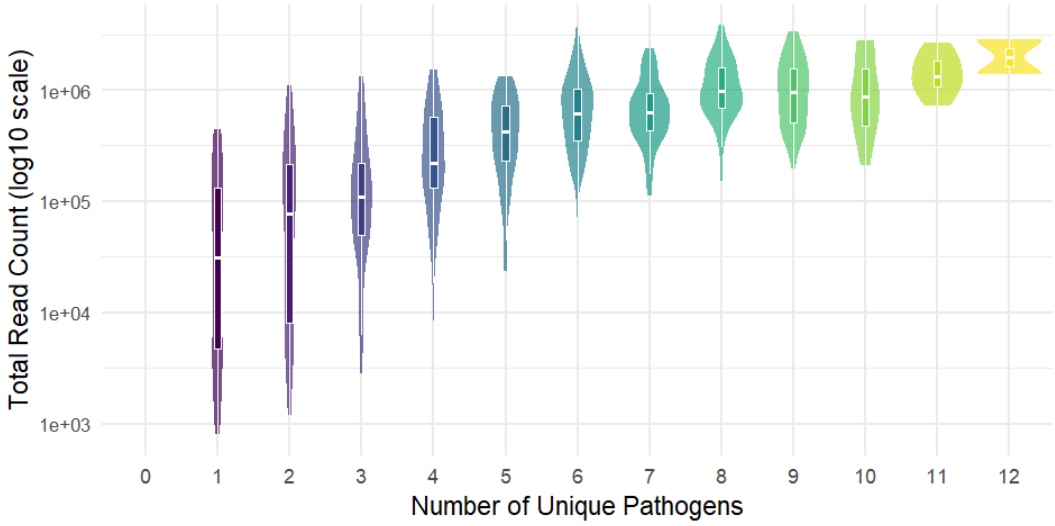


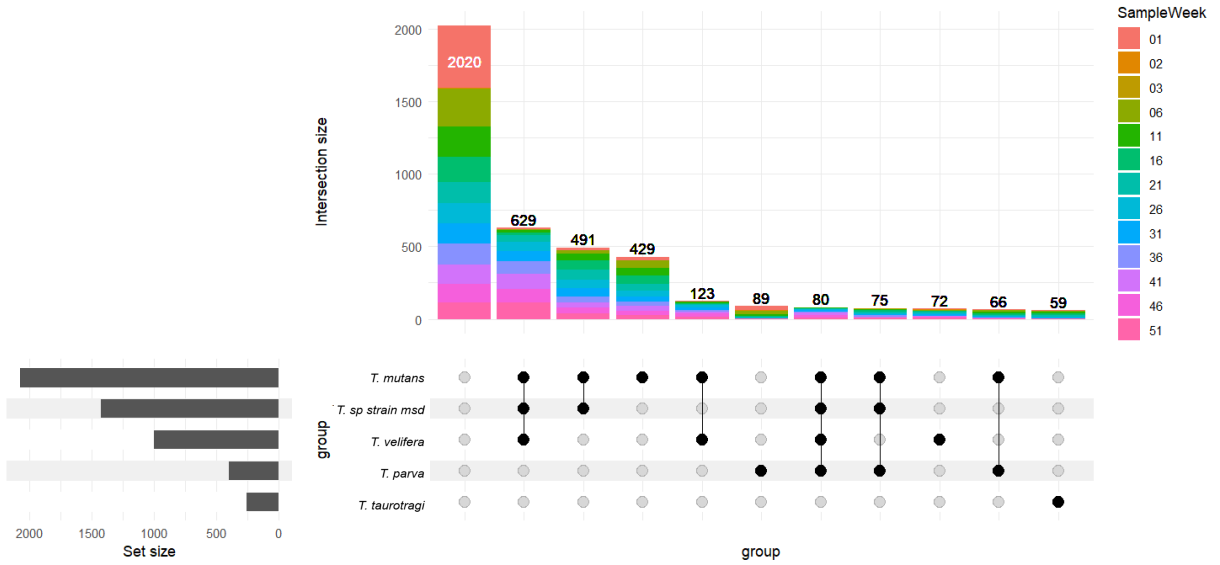
Figure 1: An analysis presented through violin plots showing the total read count associated with calves experiencing co-infections. *The y axis shows the number of unique pathogens and therefore the number of pathogens taking part in a co-infection. The X-axis shows the total read count associated with each unique pathogen co-infection count. Read count rises as co-infection count rises.*

Table 2: Frequency and mortality rates of unique haemopathogenic co-infection combinations detected across the calf cohort. *Combinations observed in fewer than six calves were grouped under "Other (rare combinations)".*

Co-infection combinations	Number of calves	Deaths	Mortality rate	Percent of total
Other (rare combinations)	362	24	6.6	76.7
None	9	3	33.3	1.9
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae	9	0	0	1.9
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay394465_mt163430_ae + theileria_mutans_af078815_tb + theileria_parva_l02366_tb + theileria_sp_strain_msd_af078816_tb + theileria_velifera_af097993_tb + uncultured_anaplasma_sp_clone_saso_ky924885_ae	9	0	0	1.9
anaplasma_platys_like_ku585990_ae	9	0	0	1.9
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay394465_mt163430_ae	8	0	0	1.7
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + theileria_mutans_af078815_tb + theileria_parva_l02366_tb + theileria_sp_strain_msd_af078816_tb + theileria_velifera_af097993_tb	8	0	0	1.7
anaplasma_platys_like_ku585990_ae + theileria_mutans_af078815_tb + theileria_sp_strain_msd_af078816_tb + theileria_velifera_af097993_tb + uncultured_anaplasma_sp_clone_saso_ky924885_ae	8	0	0	1.7
anaplasma_platys_like_ku585990_ae + theileria_mutans_af078815_tb	7	0	0	1.5
anaplasma_platys_like_ku585990_ae + uncultured_anaplasma_sp_clone_saso_ky924885_ae	7	0	0	1.5
anaplasma_bovis_ab983439_ae + anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay394465_mt163430_ae + theileria_mutans_af078815_tb +	6	0	0	1.3

theileria_parva_l02366_tb + theileria_sp_strain_msd_af078816_tb + theileria_taurotragi_l19082_tb + theileria_velifera_af097993_tb + uncultured_anaplasma_sp_clone_saso_ky924885_ae				
anaplasma_bovis_ab983439_ae + anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + theileria_mutans_af078815_tb + theileria_parva_l02366_tb + theileria_sp_strain_msd_af078816_tb + theileria_taurotragi_l19082_tb + theileria_velifera_af097993_tb + uncultured_anaplasma_sp_clone_saso_ky924885_ae	6	0	0	1.3
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay394465_mt163430_ae + theileria_mutans_af078815_tb + theileria_parva_l02366_tb + theileria_sp_strain_msd_af078816_tb + theileria_taurotragi_l19082_tb + theileria_velifera_af097993_tb + uncultured_anaplasma_sp_clone_saso_ky924885_ae	6	0	0	1.3
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay394465_mt163430_ae + theileria_mutans_af078815_tb + theileria_parva_l02366_tb + theileria_sp_strain_msd_af078816_tb + theileria_velifera_af097993_tb	6	0	0	1.3
anaplasma_bovis_u03775_ae + anaplasma_platys_like_ku585990_ae + ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay394465_mt163430_ae + theileria_mutans_af078815_tb + theileria_sp_strain_msd_af078816_tb + theileria_velifera_af097993_tb + uncultured_anaplasma_sp_clone_saso_ky924885_ae	6	0	0	1.3
theileria_parva_l02366_tb	6	5	83.3	1.3

A



B

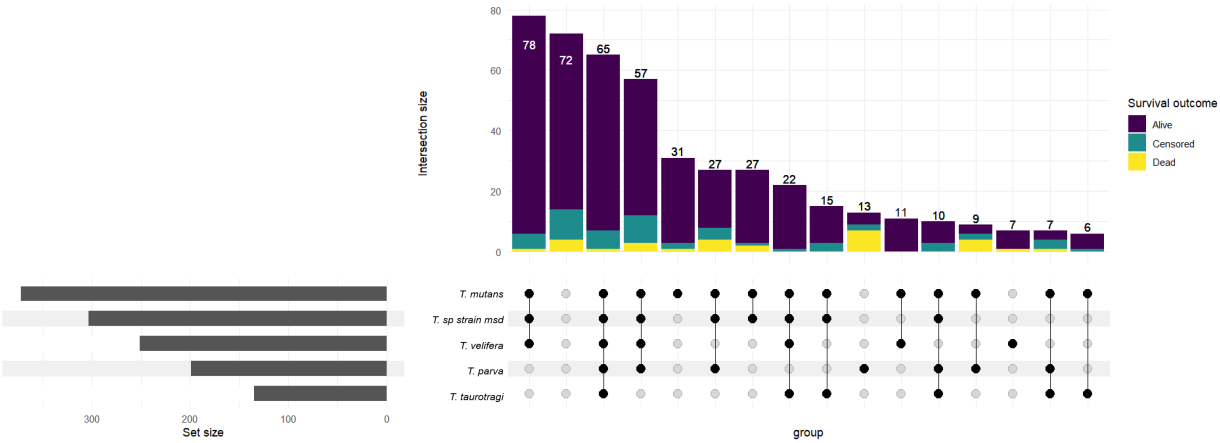
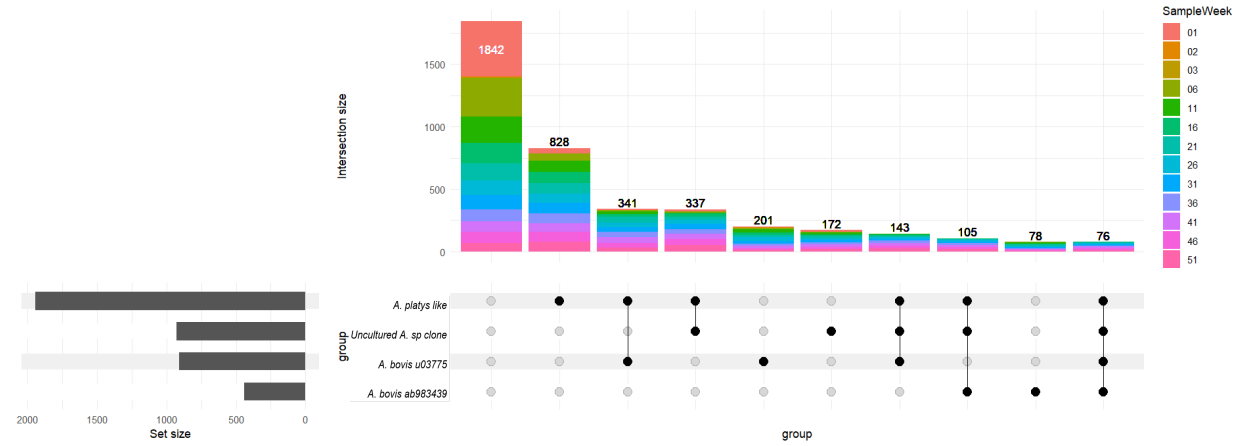


Figure 2: UpSet plot showing co-infection patterns among Theileria species detected in calves. *(A) Each vertical bar represents the number of individual samples taken infected with the corresponding combination of Theileria species shown on the x axis, this intersection matrix displays unique infection profiles. Horizontal bars show total infections per strain. Bar colour indicates the sample week in which that particular sample was taken. (B) Each vertical bar represents the calves infected with the corresponding combination of Theileria species, while horizontal bars show total infections per strain. Bar colour indicates the definitive outcome of the calf (alive, dead or censored), allowing visual comparison of co-infection profiles associated with survival or mortality.*

A



B

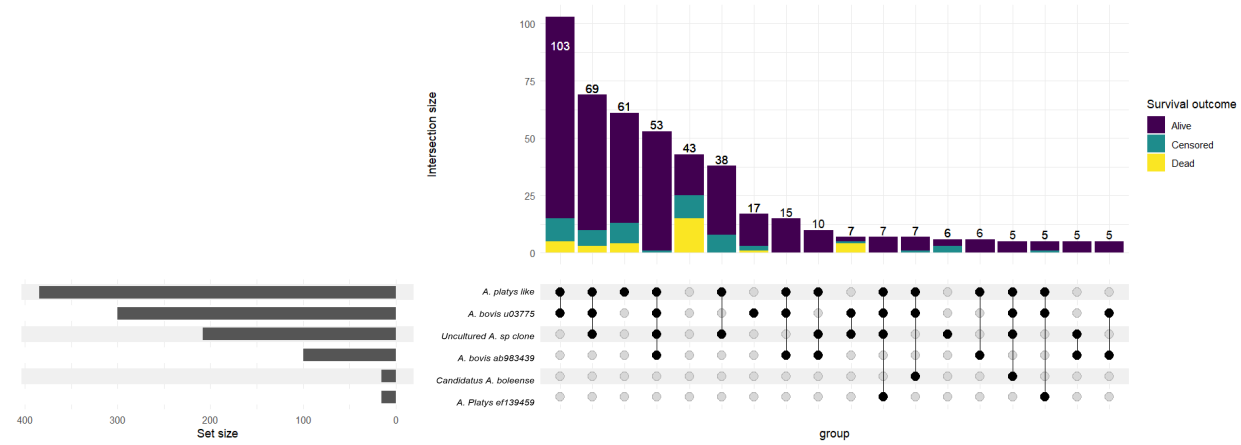


Figure 3: UpSet plot showing co-infection patterns among Anaplasma species detected in calves. *(A) Each vertical bar represents the number of individual samples taken infected with the corresponding combination of Anaplasma species shown on the x axis, this intersection matrix displays unique infection profiles. Horizontal bars show total infections per strain. Bar colour indicates the sample week in which that particular sample was taken. It only shows co-infection combinations with more than 10 occurrences. (B) Each vertical bar represents the calves infected with the corresponding combination of Anaplasma species, while horizontal bars show total infections per strain. Bar colour indicates the definitive outcome of the calf (alive, dead or censored), allowing visual comparison of co-infection profiles associated with survival or mortality.*

Table 3: Association between mortality from ECF outcome and Theileria parva infection and co-infection status. *This table shows the number of calves, deaths and mortality rates associated with T. parva infection status and the presence or absence of co-infections with other haemopathogens. Calves were grouped into three mutually exclusive categories: those infected with T. parva alone, those infected with T. parva and at least one additional pathogen, and those uninfected with T. parva.*

Infection Group	Number of Calves	Deaths	Mortality Rate (%)
T. parva only	6	5	83.3
T. parva + co-infection	193	17	8.8
No *T. parva*	273	10	3.7

Table 4: Association between mortality by ECF and exposure order status throughout the cohort.

Exposure Order	Alive	Died of ECF	Total
No exposure (3 species)	104 (95%)	6 (5.5%)	110 (100%)
T. parva infection first	286 (97%)	10 (3.4%)	296 (100%)
T. velifera/ T. mutans infection first	50 (76%)	16 (24%)	66 (100%)
Total	440 (93%)	32 (6.8%)	472 (100%)

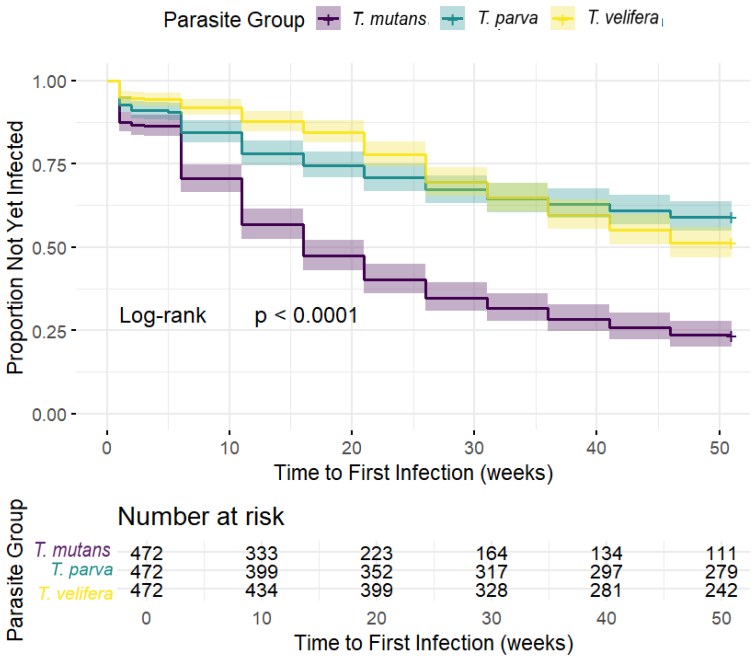


Figure 4: This Kaplan Meier plot shows the probability of remaining uninfected by the 3 Theileria species as the 51 weeks go on. *The proportion of calves becoming infected with Theileria mutans grows the fastest and remains much higher than that of Theileria parva and velifera that have similar infection patterns through time.*

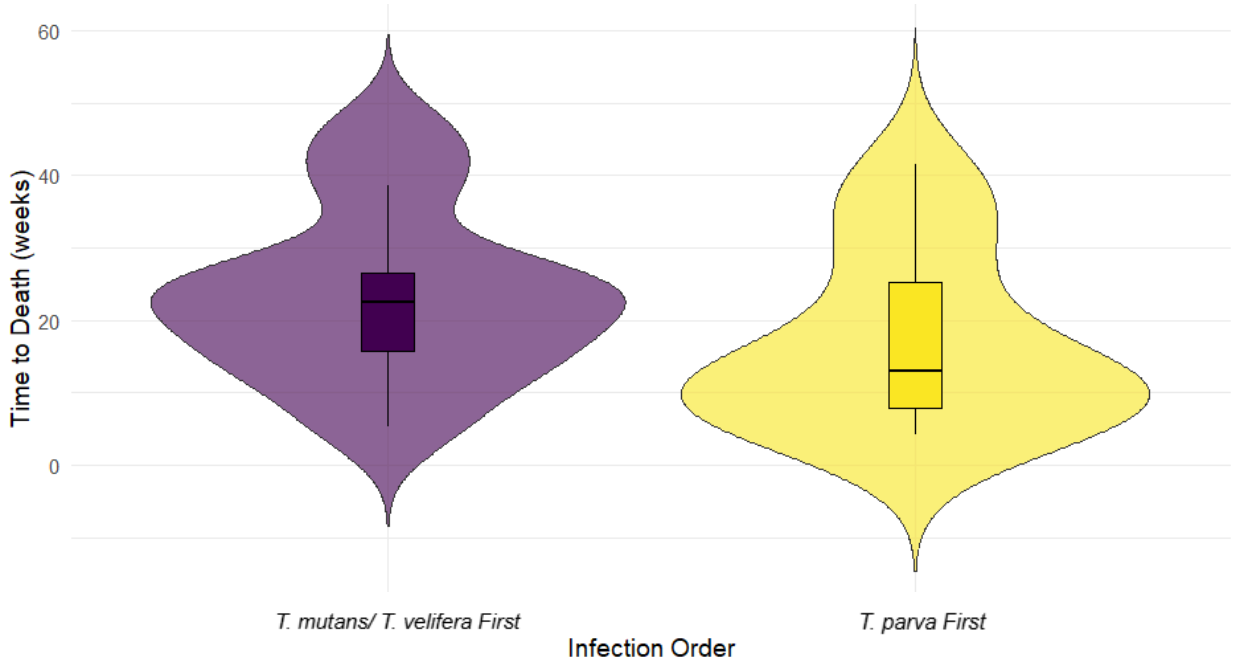


Figure 5: A violin plot showing the difference in time to death between the calves that experienced a T. mutans and/or T. velifera infection (purple) before a T. parva infection, or those that experienced a T. parva infection first (yellow). *In the V group the median time to death happened later than those in the T. parva first group. Although a Wilcoxon rank sum test did not find this statistically significant (p= 0.1471)*

Appendix C - Associations between genetics and disease outcome

Table 1: Association between genotype and survival outcome. *Table showing amount and percentages of calves in the study that died of ECF, survived the study or were censored from the study and which FAF1B genotype they were associated with.*

	Survival outcome			
	ECF	Alive	Censored	Total
Genotype				
CC	16 (6%)	225 (85%)	24 (9.1%)	265 (100%)
CT	16 (7%)	186 (82%)	25 (11%)	227 (100%)
TT	0 (0%)	49 (88%)	7 (13%)	56 (100%)
Total	32 (5.8%)	460 (84%)	56 (10%)	548 (100%)

Table 2: Association between genotype and T. parva infection in calves. *Reduced study group to those calves for which we have the haemobiome data, 472 calves in total. The table shows the number of calves with each genotype, the number and percentage of calves in each genotype infected with T. parva at some point throughout the study period.*

Genotype	No. calves	No. infected
CC	232	103 (44.4%)
CT	192	77 (40.1%)
TT	48	19 (39.6%)

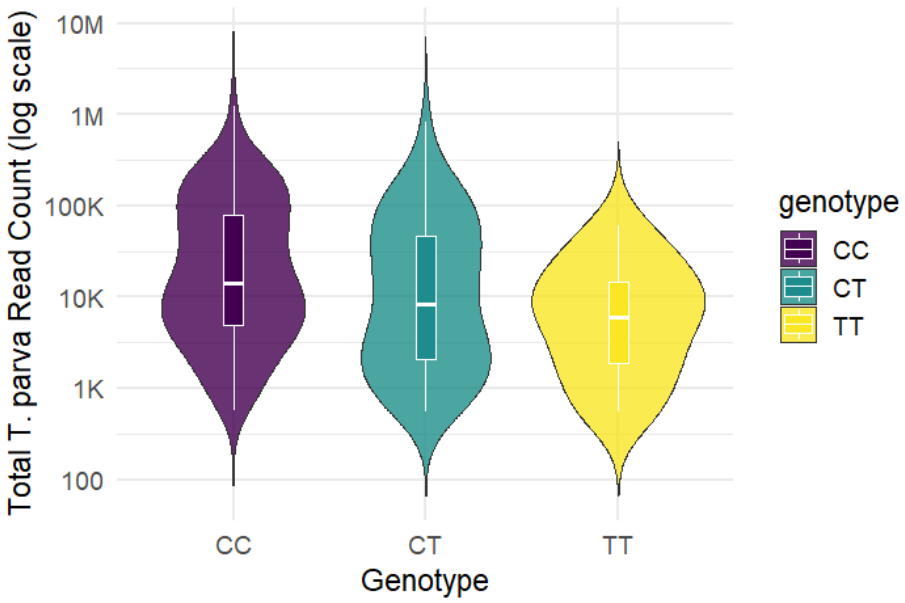


Figure 1: Violin plot analysis of total T. parva pathogen load across genotypes CC (red), CT (green) and TT (blue) of the FAF1B gene. *Genotype is shown along the x axis while pathogen load as read count is shown along the y axis. The CC genotype displayed the highest median read count, with the majority of samples clustered between 10^3 and 10^5 reads. Contrastingly, the CT and TT genotypes indicated lower median values, with a higher concentration of samples falling between 10^2 and 10^4 reads. Both CC and CT demonstrated a right-skewed distribution, indicating the presence of a few samples with exceptionally high T. parva read counts.*

Appendix D - R code

Code to organise all data into 1 data frame

```
# Libraries
library(dplyr)
library(tidyr)
library(readxl)
library(janitor)
library(tidyverse)
library(here)
library(ggplot2)
library(survival)
library(survminer)
library(scales)
library(coxphf)
library(viridis)
library(ggpubr)
library(ggsci)
library(ComplexUpset)
library(gtsummary)
library(boot)
library(vegan)
library(patchwork)
library(purrr)

################### Rearrange the miseq data
# Upload file path for the combined miseq results page
file_path_combined <- here("Edited original data", "combined results
page.xlsx")
sheet_names <- excel_sheets(file_path_combined)

# Read all miseq sheets into a vertical list
sheets_list <- lapply(sheet_names, function(sheet)
{read_excel(file_path_combined, sheet = sheet)})

# Turn combined sheets into one data frame
combined_data <- bind_rows(sheets_list)
combined_data <- combined_data %>% clean_names()

# Organize sampleID names into adequate form for analysis. By
combining rows of same sampleID. 'AE' and 'TB' results from the same
sample are now a single row for the whole sample.

# Filter for rows with `AE` or `TB`
```

```
data_ae <- combined_data %>% filter(grepl("AE", sample_id))
data_tb <- combined_data %>% filter(grepl("TB", sample_id))

# Select specific columns: column 1, and columns 6 to 23
data_ae <- data_ae %>% select(sample_id, 1: 6, 7:23, 25:27, 29:31,
33)
data_tb <- data_tb %>% select(sample_id, 1: 6, 7:23, 25:27, 29:31,
33)

# Remove the suffix "AE" or "TB" from the SampleID
data_ae <- data_ae %>% mutate(sample_id = gsub("AE", "", sample_id))
data_tb <- data_tb %>% mutate(sample_id = gsub("TB", "", sample_id))

# Join the data frames by `SampleID` to combine `AE` and `TB` data
into one row per sample
combined_data <- full_join(data_ae, data_tb, by = "sample_id", suffix
= c("_AE", "_TB"))

# Further refine combined data to only pathogens of interest, and
check for NAs
combined_data <- combined_data %>% select(2:5, 31:34,
"sample_id","anaplasma_bovis_u03775_AE",
"anaplasma_bovis_ab983439_AE", "anaplasma_marginale_cp000030_AE",
"anaplasma_platys_like_ku585990_AE",
"anaplasma_phagocytophilum_u02521_AE",
"candidatus_anaplasma_boleense_ku586025_AE",
"uncultured_anaplasma_sp_clone_saso_ky924885_AE","uncultured_anaplasm
a_sp_jn862825_AE","ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minas
ensis_af414399_ay394465_mt163430_AE",
"ehrlichia_ruminantium_x61659_AE","anaplasma_platys_ef139459_AE",
"babesia_bigemina_ay603402_TB","babesia_bigemina_lk391709_TB",
"babesia_bigemina_ku206291_TB", "theileria_mutans_af078815_TB",
"theileria_sp_strain_msd_af078816_TB","theileria_parva_l02366_TB",
"theileria_taurotragi_l19082_TB" ,
"theileria_velifera_af097993_TB","babesia_bovis_kf928959_TB","babesia
_bovis_aaxt01000002_TB",
"babesia_bovis_ay603398_TB","babesia_bigemina_lk391709_2_TB","babesia
_bovis_jq437260_TB")
combined_data[sapply(combined_data, is.numeric)] <-
  lapply(combined_data[sapply(combined_data, is.numeric)],
function(x) replace(x, is.na(x), 0))

# Trim `Sample ID` to the first 9 characters
combined_data <- combined_data %>% mutate(sample_id =
substr(sample_id, 1, 9))
```

```
########################## Combine Miseq data with Calf IDs

# Upload file path for the combined miseq results page
file_path_sample <- here("Edited original data", "ideal_sample.xlsx")
data <- read_excel(file_path_sample)


filtered_data <- data %>%
  filter(
    grepl("^RED", `SampleID`),
    grepl("RED", `Type of sample stored`)
  )

# Create a new data frame with VisitID (calf ID) and the
corresponding "IDEAL SAMPLES ID" (code)
calf_codes <- filtered_data %>%
  select(VisitID, codes = `SampleID`)

# View the result
print(calf_codes)

# Merge the calf codes dataframe to correspond with the miseq data
frame based on Sample ID so that visit ID can be seen on miseq data
merged_data <- merge(combined_data, calf_codes, by.x = "sample_id",
by.y = "codes", all = TRUE)
merged_data <- merged_data[, c("VisitID", setdiff(names(merged_data),
"VisitID"))] #rearranges so VisitID is first

############################## Add important data from ideal calf
data to database

# Upload the file path for the IDEAL calf data
file_path <- here("Edited original data", "ideal_calf.xlsx")
ideal_calf <- read_excel(file_path)

#Ensure dates are correct
# Ensure 'Visit date' is numeric
ideal_calf$`Visit date` <- as.numeric(ideal_calf$`Visit date`)
# Convert numeric Excel serial dates to Date format
ideal_calf$`Visit date` <- as.Date(ideal_calf$`Visit date`, origin =
"1899-12-30")
# Ensure 'Date last visit with data' is numeric
ideal_calf$`Date last visit with data` <- as.numeric(ideal_calf$`Date
last visit with data`)
```

```
# Convert numeric Excel serial dates to Date format
ideal_calf$`Date last visit with data` <- as.Date(ideal_calf$`Date
last visit with data`, origin = "1899-12-30")

# Combine merged_data with ideal_calf based on 'calfID' (merged_data)
and 'visitID' (ideal_calf)
final_miseq_data <- merge(merged_data, ideal_calf, by.x = "VisitID",
by.y = "VisitID", all.x = TRUE, all.y = FALSE)
#Clean data
final_miseq_data_clean <- final_miseq_data %>% clean_names()


########## Add in serology
# Upload file path for the combined miseq results page
file_path_serology <- here("Edited original data",
"Serology_data.xlsx")
serology <- read_excel(file_path_serology)
serology_clean <- serology %>% clean_names()

# Filter to rows where the contents of the "test" column begins with
"Serology"
filtered_serology <- serology_clean %>%
  filter(grepl("Serology", test))

#Filter to main columns of interest
filtered_serology_main <- filtered_serology %>%
  select(visit_id, visit_date, test, quantitative_result_number)

#Flips data frame to be horizontal not vertical
wide_serology <- filtered_serology_main %>%
  pivot_wider(
    names_from = test,
    values_from = quantitative_result_number
  )

# Clean up names
wide_serology_clean <- wide_serology %>% clean_names()


final_miseq_data_clean <- final_miseq_data_clean %>%
  left_join(
    wide_serology_clean %>%
      select(visit_id, serology_t_parva, serology_t_mutans,
serology_b_bigemina, serology_a_marginale),  # select only specific
columns you want to add
```

```
    by = "visit_id"
  )

# Reduce to no VRDs.
final_miseq_data_clean <- final_miseq_data_clean %>%
  filter(grepl("^VRC|^VCC", visit_id))

# Sample week based on first two numbers of VisitID
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(
    sample_week = case_when(
      str_starts(visit_id, "VRC") ~ as.numeric(substr(visit_id, 4,
5)),
      str_starts(visit_id, "VCC") ~ as.numeric(substr(visit_id, 4,
5)),
        #as.numeric(difftime(visit_date, date_of_birth, units =
"days")) / 7,
      TRUE ~ NA_real_  # in case it's neither VRC nor VCC
    )
  )

#Manually set error
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(sample_week = case_when(
    sample_id %in% c("RED002592", "RED003842") ~ as.numeric("5"),  #
Set specific week
    TRUE ~ sample_week
  ))

################################# Add in date of death with postportem
data

# Upload the file path for the IDEAL postmortem data
postmortem <- here("Edited original data", "ideal_postmortem.xlsx")
ideal_postmortem <- read_excel(postmortem)

ideal_postmortem_clean <- ideal_postmortem %>% clean_names()
ideal_postmortem_clean$date_of_death <-
as.numeric(ideal_postmortem_clean$date_of_death)
ideal_postmortem_clean$date_of_death <-
as.Date(ideal_postmortem_clean$date_of_death, origin = "1899-12-30")

final_miseq_data_clean <- final_miseq_data_clean %>%
  left_join(select(ideal_postmortem_clean, calf_id, date_of_death,
euthanised, immediate_cause_pathology, definitive_aetiological_cause,
```

```
definitive_cause_pathogen, contributing_pathology_1,
contributing_cause_1), by = "calf_id")

################################## Sample week - changing all the
dates to weeks of life

# Ensure all dates in date format
final_miseq_data_clean[["visit_date"]] <-
as.Date(final_miseq_data_clean[["visit_date"]], format = "%Y-%m-%d")
final_miseq_data_clean[["date_of_birth"]] <-
as.Date(final_miseq_data_clean[["date_of_birth"]], format =
"%Y-%m-%d")

# Remove rows where anaplasma_bovis_u03775_ae is NA, as it means no
miseq data
final_miseq_data_clean <- final_miseq_data_clean %>%
  filter(!is.na(anaplasma_bovis_u03775_ae))

final_miseq_data_clean <- final_miseq_data_clean %>%
  select(1:36, 40, 44, 48:49, 138:149)

# mutate data columns as they are "unknown"
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(across(c(visit_date, date_of_birth,
date_last_visit_with_data, date_of_death), ~ as.Date(.x, format =
"%Y-%m-%d")))

num_distinct_calves <- final_miseq_data_clean %>%
  summarise(n_distinct_calf = n_distinct(calf_id)) %>%
  pull(n_distinct_calf)
print(num_distinct_calves)

########################### ADD in Sublocation zones
# upload file path of sublocation data through farm data from IDEAL
sublocation <- here("Edited original data", "ideal_farm.xlsx")
sublocation_data <- read_excel(sublocation)

sublocation_data_clean <- sublocation_data %>% clean_names()


# Merge Sublocation data
final_miseq_data_clean <- final_miseq_data_clean %>%
  left_join(sublocation_data_clean %>% select(calf_id, sublocation),
by = "calf_id")
```

```
final_miseq_data_clean <- final_miseq_data_clean %>%
rename(agro_ecological_zones = sublocation)

# Clean up survival status column
final_miseq_data_clean$dead_or_alive_at_end_of_study <-
as.factor(final_miseq_data_clean$dead_or_alive_at_end_of_study)
final_miseq_data_clean$definitive_aetiological_cause <-
as.factor(final_miseq_data_clean$definitive_aetiological_cause)

#Add a haemonchosis present column
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(haemonchosis_coinfection = ifelse(contributing_cause_1 ==
"Haemonchosis", "present", "absent"))


final_miseq_data_clean$dead_or_alive_simple <- case_when(
  final_miseq_data_clean$dead_or_alive_at_end_of_study %in% c("Dead:
Infectious death", "Dead: Death by trauma") ~ "Dead",
  final_miseq_data_clean$dead_or_alive_at_end_of_study == "Alive" ~
"Alive",
  final_miseq_data_clean$dead_or_alive_at_end_of_study == "Censored"
~ "Censored",
  TRUE ~ NA_character_
)

################ Label definitive aetiological cause column
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(definitive_aetiological_cause = case_when(
    definitive_aetiological_cause == "East coast fever" ~ "Dead",
    is.na(definitive_aetiological_cause) ~ "Alive",
    definitive_aetiological_cause %in% c(
      "Haemonchosis", "Unknown", "Foreign body", "Actiomyces
pyogenes", "Trauma",
      "Heartwater", "Trypanosomiasis", "Turning sickness", "Cassava",
"Mis-mothering",
      "Bacterial pneumonia", "Black quarter", "Viral pneumonia",
"Rabies",
      "Arcanobacterium", "Babesiosis", "Salmonellosis"
    ) ~ "Censored",
    TRUE ~ definitive_aetiological_cause  # Keep others as they are
  ))

# Assign numeric event status (1 = Dead, 0 = Censored/Alive)
```

```
final_miseq_data_clean$event <-
ifelse(final_miseq_data_clean$definitive_aetiological_cause ==
"Dead", 1, 0)

# Manually set one error
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(date_last_visit_with_data = case_when(
    calf_id == "CA020610160" ~ as.Date("2008-07-17"),  # Set specific
date for this calf
    TRUE ~ date_last_visit_with_data  # Keep existing values for
others
  ))

# Calculate survival time
# Ensure survival time is numeric
final_miseq_data_clean$date_last_visit_with_data <-
as.Date(final_miseq_data_clean$date_last_visit_with_data)
final_miseq_data_clean$date_of_birth <-
as.Date(final_miseq_data_clean$date_of_birth)

# Now subtract date of death - date of birth
final_miseq_data_clean$date_of_death <-
as.Date(final_miseq_data_clean$date_of_death)
final_miseq_data_clean$time_to_event <-
as.numeric(final_miseq_data_clean$date_of_death -
final_miseq_data_clean$date_of_birth)

# Safer version preserving Date type
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(
    date_of_death = coalesce(date_of_death,
date_last_visit_with_data)
  )

# Handle the data without date of death, do it as date of last data -
date of birth
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(time_to_event = ifelse(
    is.na(date_of_death) & dead_or_alive_at_end_of_study == "Dead",
    as.numeric(date_last_visit_with_data - date_of_birth),
    as.numeric(date_of_death - date_of_birth)
  ))
final_miseq_data_clean$time_to_event <-
final_miseq_data_clean$time_to_event / 7
```

```
# Update the 'time_to_event' for alive calves (event == 0) to the
max_week
max_week <- 51
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(
    time_to_event = ifelse(dead_or_alive_at_end_of_study == "Alive",
max_week, time_to_event)  # Set to max_week for alive calves
  )

# Adjust time_to_event only for censored cases
final_miseq_data_clean <- final_miseq_data_clean %>%
  mutate(
    time_to_event = ifelse(
      dead_or_alive_at_end_of_study == "Censored",  # Only modify
censored calves
      as.numeric(date_last_visit_with_data - date_of_birth) / 7,  #
Time until last visit
      time_to_event  # Keep existing values for Alive & Dead calves
    )
  )

final_miseq_data_clean <- final_miseq_data_clean %>%
  group_by(sample_id) %>%
  summarise(across(everything(), ~ {
    non_missing_non_zero <- .x[!is.na(.x) & .x != 0]
    if (length(non_missing_non_zero) > 0) {
      non_missing_non_zero[1]
    } else {
      # fallback: use 0 if available, else NA
      fallback <- .x[!is.na(.x)]
      if (length(fallback) > 0) fallback[1] else NA
    }
  }), .groups = "drop")
```

Code for Table 1

```
# survival data select desired columns
survival_data <- final_miseq_data_clean %>%
  distinct(calf_id, .keep_all = TRUE) %>%
  select(calf_id, event, time_to_event,
definitive_aetiological_cause)

# Ensure time_to_event is numeric
```

```
survival_data$time_to_event <-
as.numeric(survival_data$time_to_event)

# Count totals
summary_stats <- survival_data %>%
  summarise(
    total_calves = n(),
    total_deaths = sum(definitive_aetiological_cause == "Dead"),
    total_alive = sum(definitive_aetiological_cause == "Alive"),
    total_censored = sum(definitive_aetiological_cause ==
"Censored"),
    median_time_to_event = median(time_to_event),
    mean_time_to_event = mean(time_to_event),
    sd_time_to_event = sd(time_to_event)
  )
```

Code for Figure 4

```
weekly_summary_stats <- survival_data %>%
  group_by(time_to_event) %>%
  summarise(
    total_calves = n(),
    total_deaths = sum(definitive_aetiological_cause == "Dead"),
    total_alive = sum(definitive_aetiological_cause == "Alive"),
    total_censored = sum(definitive_aetiological_cause ==
"Censored"),
    .groups = "drop"
  )

# Ensure definitive_aetiological_cause has a fixed factor order (to
control consistent colours)
survival_data$definitive_aetiological_cause <-
factor(survival_data$definitive_aetiological_cause,
                                       levels = c("Dead", "Censored",
"Alive"))

survival_data <- survival_data %>%
  mutate(
    event_group = ifelse(time_to_event == 51, "Week 51", "Before Week
51")
  )
```

```
ggplot(survival_data, aes(x = time_to_event, fill =
definitive_aetiological_cause)) +
  geom_histogram(binwidth = 2, color = "black") +
  facet_wrap(~event_group, scales = "free_y") +
  scale_fill_viridis_d(option = "D", begin = 0.1, end = 0.9,
direction = -1, name = "Outcome") +
  labs(
    title = "Calf Time to Event Distribution by Outcome (Excluding
Weeks 2 & 3)",
    x = "Weeks of Life",
    y = "Number of Calves"
  ) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "right")
```

Code for Figure 5

```
# Group haemopathogen markers by genus
theileria_cols <- c(
  "theileria_parva_l02366_tb",
  "theileria_mutans_af078815_tb",
  "theileria_sp_strain_msd_af078816_tb",
  "theileria_taurotragi_l19082_tb",
  "theileria_velifera_af097993_tb"
)


anaplasma_cols <- c(
  "anaplasma_bovis_u03775_ae"    ,
  "anaplasma_bovis_ab983439_ae",
  "anaplasma_marginale_cp000030_ae",
  "anaplasma_platys_like_ku585990_ae",
  "anaplasma_phagocytophilum_u02521_ae",
  "candidatus_anaplasma_boleense_ku586025_ae"        ,
  "uncultured_anaplasma_sp_clone_saso_ky924885_ae",
  "uncultured_anaplasma_sp_jn862825_ae",
  "anaplasma_platys_ef139459_ae"
)

ehrlichia_cols <- c(

"ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay3
94465_mt163430_ae"  ,
```

```
  "ehrlichia_ruminantium_x61659_ae"
)

babesia_cols <- c(
  "babesia_bigemina_ay603402_tb"  ,
  "babesia_bigemina_lk391709_tb"  ,
  "babesia_bigemina_ku206291_tb"  ,
  "babesia_bovis_kf928959_tb"     ,
  "babesia_bovis_aaxt01000002_tb" ,
  "babesia_bovis_ay603398_tb"     ,
  "babesia_bovis_jq437260_tb")

# Define the haemopathogen columns
haemopathogen_cols <- c(
  "theileria_mutans_af078815_tb",
  "theileria_sp_strain_msd_af078816_tb",
  "theileria_parva_l02366_tb",
  "theileria_velifera_af097993_tb",
  "anaplasma_bovis_ab983439_ae",
  "anaplasma_phagocytophilum_u02521_ae",
  "anaplasma_platys_like_ku585990_ae",
  "uncultured_anaplasma_sp_clone_saso_ky924885_ae"
)

pathogen_cols <- c("anaplasma_bovis_u03775_ae",
                   "anaplasma_bovis_ab983439_ae",
                   "anaplasma_marginale_cp000030_ae",
                   "anaplasma_platys_like_ku585990_ae",
                   "anaplasma_phagocytophilum_u02521_ae",
                   "candidatus_anaplasma_boleense_ku586025_ae",
                   "uncultured_anaplasma_sp_clone_saso_ky924885_ae",
                   "uncultured_anaplasma_sp_jn862825_ae",

"ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay3
94465_mt163430_ae",
                   "ehrlichia_ruminantium_x61659_ae",
                   "anaplasma_platys_ef139459_ae",
                   "babesia_bigemina_ay603402_tb",
                   "babesia_bigemina_lk391709_tb",
                   "babesia_bigemina_ku206291_tb",
                   "theileria_mutans_af078815_tb",
                   "theileria_sp_strain_msd_af078816_tb",
                   "theileria_parva_l02366_tb",
                   "theileria_taurotragi_l19082_tb",
                   "theileria_velifera_af097993_tb",
```

```
                    "babesia_bovis_kf928959_tb",
                    "babesia_bovis_aaxt01000002_tb",
                    "babesia_bovis_ay603398_tb",
                    "babesia_bigemina_lk391709_2_tb",
                    "babesia_bovis_jq437260_tb")

#Some upset graph coding
binary_pathogen_data <- final_miseq_data_clean #%>%

# in columns 3-26 change all numbers above 0 to 1
binary_pathogen_data <- binary_pathogen_data %>%
  mutate(across(11:34, ~ ifelse(. > 0, 1, 0)))

binary_infection_data <- final_miseq_data_clean %>%
  mutate(across(all_of(pathogen_cols), ~ ifelse(. > 0, 1, 0)))

# Sample week based on first two numbers of VisitID
binary_pathogen_data$SampleWeek <-
substr(binary_pathogen_data$visit_id, 4, 5)

# Ensure sample_week is a factor (important for grouping colors)
binary_pathogen_data$sample_week <-
as.factor(binary_pathogen_data$SampleWeek)


# Function to calculate weekly proportions for any haemopathogen
group
calc_prop_by_week <- function(data, cols, group_name) {
  data %>%
    group_by(sample_week, calf_id) %>%
    summarise(infected = as.integer(any(across(all_of(cols), ~ . >
0))), .groups = "drop") %>%
    group_by(sample_week) %>%
    summarise(
      group = group_name,
      n_infected = sum(infected),
      n_calves = n_distinct(calf_id),
      prop = n_infected / n_calves,
      .groups = "drop"
    )
}

# Compute proportions for each haemopathogen genus
prop_theileria <- calc_prop_by_week(binary_pathogen_data,
theileria_cols, "Theileria")
```

```
prop_theileria$sample_week =
as.numeric(as.character(prop_theileria$sample_week))
prop_theileria <- prop_theileria %>%
  filter(!sample_week %in% c(2, 3))

prop_anaplasma <- calc_prop_by_week(binary_pathogen_data,
anaplasma_cols, "Anaplasma")
prop_anaplasma$sample_week =
as.numeric(as.character(prop_anaplasma$sample_week))
prop_anaplasma <- prop_anaplasma %>%
  filter(!sample_week %in% c(2, 3))

prop_ehrlichia <- calc_prop_by_week(binary_pathogen_data,
ehrlichia_cols, "Ehrlichia")
prop_ehrlichia$sample_week =
as.numeric(as.character(prop_ehrlichia$sample_week))
prop_ehrlichia <- prop_ehrlichia %>%
  filter(!sample_week %in% c(2, 3))

prop_babesia   <- calc_prop_by_week(binary_pathogen_data,
babesia_cols, "Babesia")
prop_babesia$sample_week =
as.numeric(as.character(prop_babesia$sample_week))
prop_babesia <- prop_babesia %>%
  filter(!sample_week %in% c(2, 3))

# Combine all genus-level summaries
prop_all <- bind_rows(prop_theileria, prop_anaplasma, prop_ehrlichia,
prop_babesia)
prop_all$sample_week <-
as.numeric(as.character(prop_all$sample_week))  # Or as.Date() if
it's a date

# Set 'group' as a factor in the order of legend
prop_all$group <- factor(prop_all$group, levels = c(
"Anaplasma","Theileria","Ehrlichia","Babesia"))

# Then plot
ggplot(prop_all, aes(x = sample_week, y = prop, color = group)) +
  geom_line(size = 1) +
  geom_point(size = 2, shape = 16) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  scale_color_viridis_d(option = "D") +
  labs(
```

```
    title = "Weekly Proportion of Calves Infected with Haemopathogen
Genera",
    x = "Weeks of Life",
    y = "Proportion Infected (%)",
    color = "Pathogen Group"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "right"
  )

# Function to compute chi-squared test for each genus
chi_squared_by_age <- function(data, cols, genus_name) {
  data %>%
    mutate(infected = as.integer(rowSums(across(all_of(cols))) > 0))
%>%
    group_by(sample_week) %>%  # replace with derived age bins if
needed
    summarise(
      infected = sum(infected),
      uninfected = n() - infected,
      .groups = "drop"
    ) -> tab

  # Convert to matrix for chisq.test
  chisq_input <- as.matrix(tab[, c("infected", "uninfected")])
  rownames(chisq_input) <- tab$sample_week

  test_result <- chisq.test(chisq_input)

  list(
    genus = genus_name,
    chisq_statistic = test_result$statistic,
    p_value = test_result$p.value,
    expected = test_result$expected,
    observed = chisq_input
  )
}

# Run tests for each genus
result_theileria <- chi_squared_by_age(binary_pathogen_data,
theileria_cols, "Theileria")
result_anaplasma <- chi_squared_by_age(binary_pathogen_data,
anaplasma_cols, "Anaplasma")
```

```
result_ehrlichia <- chi_squared_by_age(binary_pathogen_data,
ehrlichia_cols, "Ehrlichia")
result_babesia   <- chi_squared_by_age(binary_pathogen_data,
babesia_cols, "Babesia")

# Combine summary results
chi_summary <- data.frame(
  Genus = c("Theileria", "Anaplasma", "Ehrlichia", "Babesia"),
  ChiSq_Statistic = c(result_theileria$chisq_statistic,
                      result_anaplasma$chisq_statistic,
                      result_ehrlichia$chisq_statistic,
                      result_babesia$chisq_statistic),
  P_Value = c(result_theileria$p_value,
              result_anaplasma$p_value,
              result_ehrlichia$p_value,
              result_babesia$p_value)
)

print(chi_summary)
```

Code for Figure 6

```
# Reshape and summarise
prop_infected_long <- binary_pathogen_data %>%
  group_by(sample_week, calf_id) %>%
  summarise(across(all_of(haemopathogen_cols), ~ as.integer(any(. >
0))), .groups = "drop") %>%
  pivot_longer(cols = all_of(haemopathogen_cols), names_to =
"pathogen", values_to = "infected") %>%
  group_by(sample_week, pathogen) %>%
  summarise(n_infected = sum(infected), n_calves =
n_distinct(calf_id), prop = n_infected / n_calves, .groups = "drop")

#Ensure sample week is numeric
prop_infected_long$sample_week =
as.numeric(as.character(prop_infected_long$sample_week))

#Remove week 2 and 3 from data
prop_infected_long <- prop_infected_long %>%
  filter(!sample_week %in% c(2, 3))

pathogen_rename <- c(
  "theileria_parva_l02366_tb" = "T. parva",
```

```
  "theileria_mutans_af078815_tb" = "T. mutans",
  "theileria_velifera_af097993_tb" = "T. velifera",
  "anaplasma_bovis_ab983439_ae"  = "A. bovis ab983439"
,
  "anaplasma_platys_like_ku585990_ae"= "A. platys like"
,
  "anaplasma_phagocytophilum_u02521_ae" = "A. phagocytophilum"
,
  "uncultured_anaplasma_sp_clone_saso_ky924885_ae" = "uncultured A.
sp clone"          ,
  "theileria_sp_strain_msd_af078816_tb"  = "T. sp strain msd")

prop_infected_long <- prop_infected_long %>%
  mutate(pathogen = recode(pathogen, !!!pathogen_rename))

# Order legend
prop_infected_long$pathogen <- factor(prop_infected_long$pathogen,
                                      levels = c("A. platys like",
"T. mutans", "T. sp strain msd",
                                                 "T. velifera",
"uncultured A. sp clone",
                                                 "A. bovis ab983439",
"T. parva", "A. phagocytophilum"))

ggplot(prop_infected_long, aes(x = sample_week, y = prop, color =
pathogen, group = pathogen)) +
  geom_line(size = 1, alpha = 0.3) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_viridis_d(option = "D") +
  labs(
    title = "Proportion of Calves Infected per Sample Week",
    x = "Weeks of Life",
    y = "Proportion Infected"
  ) +
  scale_y_continuous(labels = scales::percent) +
  theme(legend.title = element_blank())

kruskal.test(prop ~ pathogen, data = prop_infected_long)
```

Code for Figure 7

```
diversity_data <- binary_pathogen_data %>%
  rowwise() %>%
```

```
  mutate(
    richness = sum(c_across(all_of(pathogen_cols)) > 0, na.rm =
TRUE),
    simpson = if (richness == 0) NA_real_ else
      diversity(c_across(all_of(pathogen_cols)), index = "simpson")
  ) %>%
  ungroup() %>%
  mutate(sample_week = as.numeric(as.character(sample_week))) %>%   #
ensure true numbers
  filter(!sample_week %in% c(2, 3, 5))

# Average across calves at each week to get the "population-level"
trend
simpson_trend <- diversity_data %>%
  group_by(sample_week) %>%
  summarise(
    mean_simpson = mean(simpson, na.rm = TRUE),
    se = sd(simpson, na.rm = TRUE)/sqrt(n())
  )

simpson_trend <- simpson_trend %>%
  mutate(sample_week = as.numeric(as.character(sample_week)))

# Plot mean Simpson index over calf age
ggplot(simpson_trend, aes(x = as.numeric(sample_week), y =
mean_simpson)) +
  geom_line(color = "darkblue", size = 1.2) +
  geom_point(color = "darkblue", size = 2) +
  geom_ribbon(aes(ymin = mean_simpson - se, ymax = mean_simpson +
se),
              alpha = 0.2, fill = "blue") +
  labs(title = "Simpson Diversity of Haemopathogen Infections with
Age",
       x = "Weeks of Life", y = "Mean Simpson Index (±SE)") +
  theme_minimal()
```

Code for Figure 8 & 9

```
# Summarise total burden per calf
burden_df <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    tparva_load = sum(theileria_parva_l02366_tb, na.rm = TRUE),
```

```
    theileria_load = sum(across(all_of(theileria_cols)), na.rm =
TRUE),
    anaplasma_load = sum(across(all_of(anaplasma_cols)), na.rm =
TRUE),
    .groups = "drop"
  ) %>%
  left_join(final_miseq_data_clean %>%
              select(calf_id, calf_sex, sample_week,
agro_ecological_zones, definitive_aetiological_cause) %>%
              distinct(),
            by = "calf_id") %>%
  mutate(
    sex = as.factor(calf_sex),
    sublocation_zone = as.factor(agro_ecological_zones)
  )%>%
  filter(!definitive_aetiological_cause %in% c("Censored"))

#Plot burden against survival for each risk factor
plot_burden <- function(data, burden_col, group_var, title, y_label,
label_map = NULL) {
  p <- ggplot(data, aes(x = {{ group_var }}, y = {{ burden_col }},
fill = {{ group_var }})) +
    geom_violin(trim = FALSE, alpha = 0.8, color = NA) +
    geom_boxplot(width = 0.1, outlier.shape = NA, color = "white")+
    labs(title = title, x = NULL, y = y_label) +
    scale_fill_viridis_d(option = "D") +
    theme_minimal(base_size = 13) +
    theme(legend.position = "none") +
    scale_y_continuous(trans = "log1p")

  if (!is.null(label_map)) {
    p <- p + scale_x_discrete(labels = label_map)
  }

  return(p)
}

p1_breaks <- c(0, 100, 1000, 10000, 50000, 100000, 250000, 500000,
1000000, 2000000, 5000000) # Added more granular breaks at lower end
p1 <- plot_burden(burden_df, tparva_load,
definitive_aetiological_cause, "A) T. parva Load by Survival
Outcome", "T. parva Load (log scale)")+
  scale_y_continuous(trans = "log1p",  breaks = p1_breaks, labels =
scales::comma_format()) +
  theme(
```

```
    axis.title.x = element_text(size = 13, face = "bold", margin =
margin(t = 10)),
    axis.title.y = element_text(size = 13, face = "bold", margin =
margin(r = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 9),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )

p2_breaks <- c(0, 5000, 10000, 50000, 100000, 250000, 500000,
1000000, 3000000) # Added more granular breaks at lower end
p2 <- plot_burden(burden_df, theileria_load,
definitive_aetiological_cause, "B) Total Theileria Load by Survival",
"Theileria Load")+
  scale_y_continuous(trans = "log1p", breaks = p2_breaks, labels =
scales::comma_format()) +
  theme(
    axis.title.x = element_text(size = 13, face = "bold", margin =
margin(t = 10)),
    axis.title.y = element_text(size = 13, face = "bold", margin =
margin(r = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 9),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )

p3_breaks <- c(0, 5000, 10000, 50000, 100000, 250000, 500000,
1000000, 2000000, 4000000) # Added more granular breaks at lower end
p3 <- plot_burden(burden_df, anaplasma_load,
definitive_aetiological_cause, "C) Total Anaplasma Load by Survival",
"Anaplasma Load")+
  scale_y_continuous(trans = "log1p", breaks = p3_breaks, labels =
scales::comma_format()) +
  theme(
    axis.title.x = element_text(size = 13, face = "bold", margin =
margin(t = 10)),
    axis.title.y = element_text(size = 13, face = "bold", margin =
margin(r = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 9),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )

p4_breaks <- c(0, 100, 1000, 5000, 10000, 50000, 100000, 250000,
1000000) # Added more granular breaks at lower end
```

```
p4 <- plot_burden(burden_df, tparva_load, sublocation_zone, "A) T.
parva Load by sublocation zone", "T. parva Load")+
  scale_y_continuous(trans = "log1p", breaks = p4_breaks, labels =
scales::comma_format()) +
  theme(
    axis.title.x = element_text(size = 13, face = "bold", margin =
margin(t = 10)),
    axis.title.y = element_text(size = 13, face = "bold", margin =
margin(r = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 9),
    axis.text.y = element_text(size = 9),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )

p5_breaks <- c(0,100, 1000, 5000, 10000, 50000, 100000, 250000,
1000000) # Added more granular breaks at lower end
p5 <- plot_burden(burden_df, tparva_load, sex, "B) T. parva Load by
Sex", "T. parva Load")+
  scale_y_continuous(trans = "log1p", breaks = p5_breaks, labels =
scales::comma_format()) +
  theme(
    axis.title.x = element_text(size = 13, face = "bold", margin =
margin(t = 10)),
    axis.title.y = element_text(size = 13, face = "bold", margin =
margin(r = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 9),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )

# Plot 6: T. parva load by week of life
p6_breaks <- c(0, 1000, 5000, 10000, 50000, 100000, 500000, 1000000,
2000000) # Added more granular breaks at lower end
p6 <- ggplot(burden_df, aes(x = as.numeric(sample_week), y =
tparva_load)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess") +
  labs(title = "C) T. parva Load Across Weeks of Life", x = "Sample
Week", y = "T. parva Load") +
  theme_minimal(base_size = 13) +
  scale_y_continuous(trans = "log1p", breaks = p6_breaks, labels =
scales::comma_format()) +
  theme(
    axis.title.x = element_text(size = 13, face = "bold", margin =
margin(t = 10)),
```

```
    axis.title.y = element_text(size = 13, face = "bold", margin =
margin(r = 10)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 9),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )

wilcox.test(tparva_load ~ definitive_aetiological_cause, data =
burden_df)
# Get medians and IQRs
burden_df %>%
  group_by(definitive_aetiological_cause) %>%
  summarise(
    median_load = median(tparva_load, na.rm = TRUE),
    IQR_low = quantile(tparva_load, 0.25, na.rm = TRUE),
    IQR_high = quantile(tparva_load, 0.75, na.rm = TRUE),
    n = n()
  )

wilcox.test(theileria_load ~ definitive_aetiological_cause, data =
burden_df)
burden_df %>%
  group_by(definitive_aetiological_cause) %>%
  summarise(
    median_load = median(theileria_load, na.rm = TRUE),
    IQR_low = quantile(theileria_load, 0.25, na.rm = TRUE),
    IQR_high = quantile(theileria_load, 0.75, na.rm = TRUE),
    n = n()
  )
wilcox.test(anaplasma_load ~ definitive_aetiological_cause, data =
burden_df)
burden_df %>%
  group_by(definitive_aetiological_cause) %>%
  summarise(
    median_load = median(anaplasma_load, na.rm = TRUE),
    IQR_low = quantile(anaplasma_load, 0.25, na.rm = TRUE),
    IQR_high = quantile(anaplasma_load, 0.75, na.rm = TRUE),
    n = n()
  )
kruskal.test(tparva_load ~ agro_ecological_zones, data = burden_df)
burden_df %>%
  group_by(agro_ecological_zones) %>%
  summarise(
    median_load = median(tparva_load, na.rm = TRUE),
    IQR_low = quantile(tparva_load, 0.25, na.rm = TRUE),
```

```
    IQR_high = quantile(tparva_load, 0.75, na.rm = TRUE),
    n = n()
  )
wilcox.test(tparva_load ~ calf_sex, data = burden_df)
burden_df %>%
  group_by(calf_sex) %>%
  summarise(
    median_load = median(tparva_load, na.rm = TRUE),
    IQR_low = quantile(tparva_load, 0.25, na.rm = TRUE),
    IQR_high = quantile(tparva_load, 0.75, na.rm = TRUE),
    n = n()
  )

# Arrange and print plots
ggarrange(p1, p2, p3, ncol = 2, nrow = 2)
ggarrange(p4, p5, ncol = 2, nrow = 2)
```

Code for Figure 10

```
# Summarize survival data grouped by calf_id
windowed_data_sex <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarize(
    calf_sex = first(calf_sex),
    time_to_event = max(time_to_event, na.rm = TRUE),
    event = max(event, na.rm = TRUE)
  ) %>%
  ungroup()

# Ensure time_to_event is numeric
windowed_data_sex$time_to_event <-
as.numeric(windowed_data_sex$time_to_event)

# Kaplan-Meier fit by sex
km_fit_sex <- survfit(Surv(time_to_event, event) ~ calf_sex, data =
windowed_data_sex)

# Set viridis colours (2 categories)
viridis_palette <- viridis(2, option = "D")

# Plot
ggsurvplot(
  km_fit_sex,
  data = windowed_data_sex,
```

```
  conf.int = TRUE,
  pval = TRUE,
  pval.method = TRUE,                # show test used (log-rank)
  pval.size = 5,                     # increase font size
  pval.coord = c(2, 0.82),          # position: x = 10 weeks, y = 0.78
  risk.table = TRUE,
  censor = TRUE,
  palette = viridis_palette,
  ggtheme = theme_minimal(base_size = 14),
  title = "Kaplan-Meier Survival Curve: Impact of Calf Sex on
Mortality",
  ylab = "Survival Probability",
  xlab = "Time to Death (Weeks)",
  ylim = c(0.75, 1)
)

# cox model
cox_model_sex <- coxph(Surv(time_to_event, event) ~ calf_sex, data =
windowed_data_sex)

# Display summary of the Cox model
summary(cox_model_sex)
```

Code for Figure 11

```
surv_sub <- final_miseq_data_clean %>%
  select(calf_id, sample_week, time_to_event,
definitive_aetiological_cause, event, agro_ecological_zones)

surv_subs <- surv_sub %>%
  group_by(calf_id) %>%
  summarise(
    time_to_event = first(time_to_event),       # use existing column
    event = first(event),                       # keep event
indicator
    agro_ecological_zones = first(agro_ecological_zones),
    .groups = "drop"
  )

survdiff(Surv(time_to_event, event) ~ agro_ecological_zones, data =
surv_subs)

# Summarise death rates by sublocation
death_summary <- surv_subs %>%
```

```
  group_by(agro_ecological_zones) %>%
  summarise(
    n = n(),
    deaths = sum(event),
    death_rate = deaths / n
  ) %>%
  arrange(desc(death_rate))

ggplot(death_summary, aes(x = reorder(agro_ecological_zones,
death_rate), y = death_rate, fill = agro_ecological_zones)) +
  geom_col(color = "black") +
  scale_fill_viridis_d(option = "D") +
  coord_flip() +
  labs(
    title = "Mortality Percentage by Sublocation Zone",
    x = "Sublocation Zone",
    y = "Mortality Percentage"
  ) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```

Code for Figure 12

```
# Identify FIRST T. parva infection per calf
first_tparva_week <- final_miseq_data_clean %>%
  filter(theileria_parva_l02366_tb > 0) %>%
  group_by(calf_id) %>%
  summarise(first_week = min(sample_week, na.rm = TRUE), .groups =
"drop")

first_tparva_week <- first_tparva_week %>%
  filter(!first_week %in% c(2, 3))

# Merge with survival info (grouped by calf)
surv_info <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    time_to_event = max(time_to_event, na.rm = TRUE),
    event = max(event, na.rm = TRUE),
    .groups = "drop"
  )

# Join survival + infection timing, exclude calves never infected
```

```
km_df <- first_tparva_week %>%
  left_join(surv_info, by = "calf_id") %>%
  filter(!is.na(first_week)) %>%  # only those infected
  mutate(
    infection_timing = ifelse(first_week <= 25, "Early Infection",
"Late Infection"),
    infection_timing = factor(infection_timing, levels = c("Early
Infection", "Late Infection"))
  )

# Fit Kaplan-Meier survival model
km_fit <- survfit(Surv(time_to_event, event) ~ infection_timing, data
= km_df)

viridis_palette <- viridis(3, option = "D")

# Plot
ggsurvplot(
  km_fit,
  data = km_df,
  conf.int = TRUE,
  risk.table = TRUE,
  pval = TRUE,
  #pval.method = TRUE,                # show test used (log-rank)
  pval.size = 5,                     # increase font size
  pval.coord = c(2, 0.25),
  censor = TRUE,
  palette = viridis_palette,
  ggtheme = theme_minimal(base_size = 14),
  title = "Survival by Timing of T. parva Infection",
  xlab = "Time to Death (Weeks)",
  ylab = "Survival Probability",
  legend.title = "Infection Timing",
  legend.labs = c("Early infection", "Late infection")
)
```

Code for Figure 13

```
# Your rename mapping
column_rename <- c(
  "T. parva" =   "theileria_parva_l02366_tb",
  "T. mutans" = "theileria_mutans_af078815_tb",
  "T. velifera" = "theileria_velifera_af097993_tb",
```

```
  "A. bovis ab983439" = "anaplasma_bovis_ab983439_ae",
  "A. platys like" = "anaplasma_platys_like_ku585990_ae",
  "A. phagocytophilum" = "anaplasma_phagocytophilum_u02521_ae",
  "uncultured A. sp clone" =
"uncultured_anaplasma_sp_clone_saso_ky924885_ae",
  "T. sp strain msd" = "theileria_sp_strain_msd_af078816_tb"
)

# Apply renaming to your dataframe
binary_pathogen_data <- binary_pathogen_data %>%
  rename(!!!column_rename)

# Update haemopathogen_cols to match the new names
haemopathogen_cols <- names(column_rename)

# By sample week
upset(
  binary_pathogen_data,
  intersect = haemopathogen_cols,
  min_size = 50,
  width_ratio = 0.3,
  base_annotations = list(
    'Intersection size' = intersection_size(
      mapping = aes(fill = SampleWeek)
    )
  )
) +
  scale_fill_viridis_d(option = "D", name = "Sample Week") +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),   # Remove bottom axis text
    axis.ticks.x = element_blank(),  # Remove tick marks
    plot.margin = margin(10, 10, 10, 10)  # Optional: add space
  )

# By survival outcome
binary_pathogen_data_per_calf <- binary_pathogen_data %>%
  group_by(calf_id) %>%
  summarise(across(all_of(haemopathogen_cols), ~ as.integer(any(. >
0))))

# Now add in aetiological cause
cause <- binary_pathogen_data %>%
  select(calf_id, definitive_aetiological_cause) %>%
  distinct()
```

```
binary_pathogen_data_per_calf_annotated <-
binary_pathogen_data_per_calf %>%
  left_join(cause, by = "calf_id")
# Apply renaming to your dataframe

binary_pathogen_data_per_calf_annotated <-
binary_pathogen_data_per_calf_annotated %>%
  rename(!!!column_rename)

# Update haemopathogen_cols to match the new names
haemopathogen_cols <- names(column_rename)

upset(
  binary_pathogen_data_per_calf_annotated,
  intersect = haemopathogen_cols,
  min_size = 5,
  width_ratio = 0.4,
  set_sizes = upset_set_size(),
  base_annotations = list(
    'Intersection size' = intersection_size(
      mapping = aes(fill = definitive_aetiological_cause)
    ) +
      scale_fill_viridis_d(name = "Survival outcome", option = "D")
  )
)
```

Code for Figure 14

```
# Identify earliest infection for each pathogens
infection_data <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarize(
    earliest_mutans =
suppressWarnings(min(sample_week[theileria_mutans_af078815_tb > 0],
na.rm = TRUE)),
    earliest_velifera =
suppressWarnings(min(sample_week[theileria_velifera_af097993_tb > 0],
na.rm = TRUE)),
    earliest_parva =
suppressWarnings(min(sample_week[theileria_parva_l02366_tb > 0],
na.rm = TRUE)),
    .groups = "drop"
  ) %>%
```

```
  mutate(across(starts_with("earliest_"), ~ ifelse(is.infinite(.),
NA_real_, .)))  # Convert Inf to NA

# Merge with main dataset
windowed_data <- final_miseq_data_clean %>%
  left_join(infection_data, by = "calf_id") %>%
  mutate(
    # Replace NA with Inf for comparison purposes
    earliest_mutans = ifelse(is.na(earliest_mutans), Inf,
earliest_mutans),
    earliest_velifera = ifelse(is.na(earliest_velifera), Inf,
earliest_velifera),
    earliest_parva = ifelse(is.na(earliest_parva), Inf,
earliest_parva),

    # Define infection order groups
    infection_order = case_when(
      earliest_mutans < earliest_parva & earliest_mutans <
earliest_velifera ~ "Mutans First",
      earliest_velifera < earliest_parva & earliest_velifera <
earliest_mutans ~ "Velifera First",
      (earliest_mutans == earliest_velifera) & (earliest_mutans <
earliest_parva) ~ "T. Mutans/T. Velifera First",
      earliest_parva < earliest_mutans & earliest_parva <
earliest_velifera ~ "T. Parva First",
      TRUE ~ "No exposure (3 strains)"
    )
  ) %>%
  # Convert Inf back to NA for clarity in the final output
  mutate(across(starts_with("earliest_"), ~ ifelse(. == Inf,
NA_real_, .))) %>%
  group_by(calf_id) %>%
  summarize(
    infection_order = first(infection_order),
    time_to_event = max(time_to_event, na.rm = TRUE),
    event = max(event, na.rm = TRUE),
    .groups = "drop"
  )

# Group Mutans First, Velifera First, and Mutans/Velifera First into
one category
windowed_data <- windowed_data %>%
  mutate(infection_order = case_when(
    infection_order %in% c("Mutans First", "Velifera First", "T.
Mutans/T. Velifera First") ~ "T. Mutans/T. Velifera First",
```

```
    TRUE ~ infection_order  # Keep other values unchanged
  ))

# Convert to factor for plotting
windowed_data$infection_order <-
factor(windowed_data$infection_order,
                                          levels = c("No exposure (3
strains)", "T. Mutans/T. Velifera First", "T. Parva First"))

windowed_data_detailed <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    tparva_load = sum(theileria_parva_l02366_tb, na.rm = TRUE),
    outcome = first(definitive_aetiological_cause),
    .groups = "drop"
  ) %>%
  left_join(windowed_data, by = "calf_id") %>%
  filter(outcome %in% c("Alive", "Dead"))

windowed_data_detailed <- windowed_data_detailed %>%
  filter(infection_order %in% c("T. Mutans/T. Velifera First", "T.
Parva First"))

ggplot(windowed_data_detailed, aes(x = infection_order, y =
tparva_load, fill = outcome)) +
  geom_violin(position = position_dodge(width = 0.9), trim = FALSE,
alpha = 0.8) +
  geom_boxplot(position = position_dodge(width = 0.9), width = 0.1,
               outlier.shape = NA, color = "black") +
  scale_y_continuous(
    trans = "log1p",
    breaks = c(1e2, 1e3, 1e4, 1e5, 1e6, 1e7),
    labels = label_scientific(digits = 1)
  ) +
  scale_fill_viridis_d(option = "D", name = "Outcome") +
  labs(
    title = "T. parva Load by Infection Order and Survival Outcome",
    x = "Infection Order",
    y = "Total T. parva Read Count (log scale)"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(size = 12),
    axis.title.x = element_text(face = "bold"),
```

```
    axis.title.y = element_text(face = "bold")
  )
```

Code for Figure 15

```
# Kaplan-Meier survival fit
km_fit <- survfit(Surv(time_to_event, event) ~ infection_order, data
= windowed_data)

# Plot Kaplan-Meier curve
ggsurvplot(km_fit, data = windowed_data,
           conf.int = TRUE,
           pval = TRUE,
           pval.method = TRUE,               # show test used
(log-rank)
           pval.size = 5,                    # increase font size
           pval.coord = c(2, 0.65),
           risk.table = TRUE,
           censor = TRUE,
           ggtheme = theme_minimal(),
           title = "Kaplan-Meier Survival by infection Order",
           ylim = c(0.6, 1),
           xlab = "Time to Death (Weeks)",
           palette = viridis_palette
)

cox_model <- coxph(Surv(time_to_event, event) ~ infection_order, data
= windowed_data)

# View model summary
summary(cox_model)
# Test proportional hazards assumption
cox.zph(cox_model)
```

Code for Figure 16

```
Genomics <- here("Edited original data", "Genomics.xlsx")
Genomics <- read_excel(Genomics)
Genomics_clean <- Genomics %>% clean_names()

# Merge the genomics and calf ID dataframe

# Merge two data frames by 'calf_id', keeping all columns from both
```

```
merged_genomics <- full_join(final_miseq_data_clean, Genomics_clean,
by = "calf_id")
merged_genomics <- merged_genomics %>%
  group_by(calf_id) %>%
  slice(1) %>%  # Select the first row of each group
  ungroup()
merged_genomics <- select(merged_genomics, calf_id, event,died,
time_to_event, genotype, definitive_aetiological_cause, event,
time_to_event)

# Manually remove errors
merged_genomics <- merged_genomics %>%
  filter(!calf_id %in% c("CA020610172", "CA051910553"))

# clean up data
merged_genomics <- merged_genomics %>%
  mutate(
    definitive_aetiological_cause =
as.character(definitive_aetiological_cause),  # Convert factor to
character
    definitive_aetiological_cause = ifelse(died == "No", "Alive",
definitive_aetiological_cause),
    event = ifelse(died == "No", 0, event),
    time_to_event = ifelse(is.na(time_to_event) & died == "No", 51,
time_to_event)  # Update only if NA
  ) %>%
  mutate(definitive_aetiological_cause =
as.factor(definitive_aetiological_cause))  # Convert back to factor
if needed

# Convert 'genotype' to a factor
merged_genomics$genotype <- as.factor(merged_genomics$genotype)

# Summarize survival data grouped by calf_id
merged_genomics <- merged_genomics %>%
  group_by(calf_id) %>%
  summarize(
    genotype = first(genotype),  # Get sex for each calf
    time_to_event = max(time_to_event, na.rm = TRUE),
    event = max(event, na.rm = TRUE)  # Retains event=0 for alive &
censored
  ) %>%
  ungroup()

# Fit Kaplan-Meier model by genotype
```

```
km_fit_genomics <- survfit(Surv(time_to_event, event) ~ genotype,
data = merged_genomics)

# Plot Kaplan-Meier curve
ggsurvplot(
  km_fit_genomics,
  data = merged_genomics,
  conf.int = TRUE,
  pval = TRUE,
  pval.method = TRUE,               # show test used (log-rank)
  pval.size = 5,                    # increase font size
  pval.coord = c(5, 0.87),
  risk.table = TRUE,
  censor = TRUE,
  censor.shape = "|",  # Show censored calves as vertical ticks
  censor.size = 3,
  palette = viridis_palette,
  ggtheme = theme_minimal(base_size = 14),
  title = "Kaplan-Meier Survival Curve: Impact of Genotype on
mortality",
  ylim = c(0.85, 1),
  xlab = "Time to Death (Weeks)"
)
```

Code for Table 2

```
merged_genomics$genotype <-
relevel(as.factor(merged_genomics$genotype), ref = "CC")

firth_model <- coxphf(Surv(time_to_event, event) ~ genotype, data =
merged_genomics)
summary(firth_model)
```

Code for Figure 17

```
# Create burden summary (one row per calf_id)
burden_summary <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    theileria_load = sum(across(all_of(theileria_cols)), na.rm =
TRUE),
    anaplasma_load = sum(across(all_of(anaplasma_cols)), na.rm =
TRUE),
```

```
    .groups = "drop"
  )

# Count unique pathogens detected per calf
co_infection_count <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    co_infection_n = sum(sapply(across(all_of(pathogen_cols)),
function(x) any(x > 0, na.rm = TRUE))),
    .groups = "drop"
  )

# T. parva load
tparva_load_df <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    tparva_load = sum(theileria_parva_l02366_tb, na.rm = TRUE),
    .groups = "drop"
  )

# Merge everything into combined_data
combined_data <- windowed_data %>%
  select(calf_id, time_to_event, event, infection_order) %>%
  left_join(Genomics_clean %>% select(calf_id, genotype), by =
"calf_id") %>%
  left_join(tparva_load_df, by = "calf_id") %>%
  left_join(burden_summary, by = "calf_id") %>%   # << added here
  left_join(co_infection_count, by = "calf_id") %>%
  filter(!is.na(event), !is.na(time_to_event), !is.na(genotype)) %>%
  mutate(
    genotype = factor(genotype),
    infection_order = factor(infection_order),
    genotype = relevel(genotype, ref = "CC"),
    infection_order = relevel(infection_order, ref = "No infection")
  )

combined_data <- combined_data %>%
  mutate(
    log_tparva    = as.numeric(scale(log1p(tparva_load))),
    log_anaplasma = as.numeric(scale(log1p(anaplasma_load))),
    log_theileria = as.numeric(scale(log1p(theileria_load)))
  )

cox_firth <- coxphf(
```

```
  Surv(time_to_event, event) ~ genotype + infection_order +
co_infection_n
  + log_tparva + log_anaplasma + log_theileria,
  data = combined_data,
  maxit = 200,
  maxstep = 0.5
)

summary(cox_firth)

# Plot it
hr_data <- data.frame(
  variable = c("Genotype: CT", "Genotype: TT",
               "Infection order: Mutans/Velifera First",
               "Infection order: Parva First",
               "Co-infection count",
               "log(T. parva load)",
               "log(Anaplasma load)",
               "log(Theileria load)"),
  HR = c(0.6391, 0.2946, 0.9193, 6.3578, 0.6625, 2.1904, 0.4743,
0.8087),
  lower_CI = c(0.2985, 0.0023, 0.2314, 1.9241, 0.4845, 1.1719,
0.3221, 0.3389),
  upper_CI = c(1.3419, 2.2669, 4.1193, 27.3617, 0.8779, 4.1294,
0.6899, 1.7924))


hr_data <- hr_data %>%
  mutate(
    p_value = case_when(
      is.na(lower_CI) ~ NA_real_,
      lower_CI > 1 | upper_CI < 1 ~ 0.01,
      TRUE ~ 0.2  # otherwise not significant
    ),
    sig = case_when(
      is.na(p_value) ~ "",
      p_value < 0.001 ~ "***",
      p_value < 0.01  ~ "**",
      p_value < 0.05  ~ "*",
      TRUE ~ ""
    )
  )

# Add a baseline reference (HR = 1) and label significance
```

```
ggplot(hr_data, aes(x = variable, y = HR, ymin = lower_CI, ymax =
upper_CI)) +
  geom_pointrange() +
  geom_hline(yintercept = 1, linetype = "dashed", color = "red") +  #
baseline
  geom_text(aes(label = sig, y = upper_CI * 1.1),  # place stars
slightly above CI
            size = 4, color = "black") +
  coord_flip() +
  scale_y_log10() +
  labs(
    x = "",
    y = "Hazard Ratio (log scale)",
    title = "Cox-Firth Regression: Hazard Ratios with 95% CI"
  ) +
  theme_minimal()
```

Code for Table 3

```
model_full <- coxphf(
  Surv(time_to_event, event) ~ genotype + infection_order +
co_infection_n + log_tparva + log_anaplasma + log_theileria,
  data = combined_data
)

# Reduced models (drop one covariate at a time)
model_no_genotype   <- update(model_full, . ~ . - genotype)
model_no_infection  <- update(model_full, . ~ . - infection_order)
model_no_coinf      <- update(model_full, . ~ . - co_infection_n)
model_no_tparva     <- update(model_full, . ~ . - log_tparva)
model_no_anaplas    <- update(model_full, . ~ . - log_anaplasma)
model_no_theil     <- update(model_full, . ~ . - log_theileria)


extract_info <- function(model, name) {
  ll <- logLik(model)
  k  <- length(coef(model))   # number of parameters
  aic <- -2 * as.numeric(ll) + 2 * k
  data.frame(Model = name, logLik = as.numeric(ll), k = k, AIC = aic)
}

results <- rbind(
  extract_info(model_full, "Full model"),
  extract_info(model_no_genotype, "No genotype"),
```

```
  extract_info(model_no_infection, "No infection order"),
  extract_info(model_no_coinf, "No co-infection count"),
  extract_info(model_no_tparva, "No Theileria parva"),
  extract_info(model_no_anaplas, "No Anaplasma"),
  extract_info(model_no_theil, "No Theileria ")
)

results$Delta_AIC <- results$AIC - min(results$AIC)

lr_test <- function(full, reduced) {
  test <- anova(full, reduced)  # works if both are coxphf
  test$"Pr(>Chi)"[2]
}

pvals <- c(
  NA,
  lr_test(model_full, model_no_genotype),
  lr_test(model_full, model_no_infection),
  lr_test(model_full, model_no_coinf),
  lr_test(model_full, model_no_tparva),
  lr_test(model_full, model_no_anaplas),
  lr_test(model_full, model_no_theil)

)

results$p_value <- pvals

results
```

Code for Appendix A Figure 1

```
###(For this only complete data wrangle up to #### Label definitive
aetiological cause column step)

# Cause of death summary
cause_of_death_summary <- final_miseq_data_clean %>%
  filter(!is.na(definitive_aetiological_cause)) %>%  # Only rows
where calf died
  distinct(calf_id, .keep_all = TRUE) %>%            # Only one entry
per calf
  count(definitive_aetiological_cause, sort = TRUE) %>%
  rename(Cause = definitive_aetiological_cause, Count = n)

# Plot it
```

```
ggplot(cause_of_death_summary, aes(x = reorder(Cause, -Count), y =
Count, fill = Cause)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Count), vjust = -0.3, size = 4) +
  scale_fill_viridis_d(option = "D") +
  labs(title = "Number of Calf Deaths by Cause",
       x = "Cause of Death",
       y = "Number of Calves") +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none",
        plot.title = element_text(face = "bold", hjust = 0.5))
```

Code for Appendix A Figure 2

```
pathogen_rename <- c(
  "theileria_parva_l02366_tb" = "T. parva",
  "theileria_mutans_af078815_tb" = "T. mutans",
  "theileria_velifera_af097993_tb" = "T. velifera",
  "anaplasma_marginale_cp000030_ae" = "A. marginale",
  "anaplasma_bovis_u03775_ae" = "A. bovis u03775",
  "ehrlichia_ruminantium_x61659_ae" = "E. ruminantium",
  "babesia_bigemina_lk391709_2_tb" = "B. bigemina lk391709 1",
  "anaplasma_bovis_ab983439_ae"  = "A. bovis ab983439",
  "anaplasma_platys_like_ku585990_ae"= "A. platys like",
  "anaplasma_phagocytophilum_u02521_ae" = "A. phagocytophilum",
  "candidatus_anaplasma_boleense_ku586025_ae" = "candidatus A.
boleense",
  "uncultured_anaplasma_sp_clone_saso_ky924885_ae" = "uncultured A.
sp clone",
  "uncultured_anaplasma_sp_jn862825_ae" = "uncultured A. sp",

"ehrlichia_sp_tibet_ehrlichia_canis_ehrlichia_minasensis_af414399_ay3
94465_mt163430_ae" = "E. sp tibet/canis/minasensis",
  "anaplasma_platys_ef139459_ae" = "A. platys",
  "babesia_bigemina_ay603402_tb" = "B. bigemina ay603402",
  "babesia_bigemina_lk391709_tb"  = "B. bigemina lk391709",
  "babesia_bigemina_ku206291_tb"   = "B. bigemina ku206291",
  "theileria_sp_strain_msd_af078816_tb"  = "T. sp strain msd",
  "theileria_taurotragi_l19082_tb" = "T. taurotragi ",
  "babesia_bovis_kf928959_tb"  = "B. bovis kf928959",
  "babesia_bovis_aaxt01000002_tb"  = "B. bovis aaxt01000002",
```

```
  "babesia_bovis_ay603398_tb"   = "B. bovis ay603398",
  "babesia_bovis_jq437260_tb" = "B. bovis jq437260")

prop_infected_long <- binary_pathogen_data %>%
  group_by(sample_week, calf_id) %>%
  summarise(across(all_of(pathogen_cols), ~ as.integer(any(. > 0))),
.groups = "drop") %>%
  pivot_longer(cols = all_of(pathogen_cols), names_to = "pathogen",
values_to = "infected") %>%

  # Add n_calves per week inline
  left_join(
    binary_pathogen_data %>%
      group_by(sample_week) %>%
      summarise(n_calves = n_distinct(calf_id), .groups = "drop"),
    by = "sample_week"
  ) %>%

  group_by(sample_week, pathogen) %>%
  summarise(n_infected = sum(infected), n_calves = first(n_calves),
.groups = "drop") %>%
  mutate(
    prop = n_infected / n_calves,
    sample_week = as.numeric(as.character(sample_week)),
    pathogen = recode(pathogen, !!!pathogen_rename)
  ) %>%
  group_by(pathogen) %>%
  filter(!sample_week %in% c(2, 3)) %>%
  filter(max(prop) >= 0.1) %>%  # remove strains that never exceed
10% prevalence
  ungroup()

# proportional stacked area chart (each layer is the share of total
infections that week)
ggplot(prop_infected_long, aes(x = sample_week, y = prop, fill =
pathogen)) +
  geom_area(position = "fill", alpha = 0.9, color = "white", size =
0.2) +
  scale_x_continuous(breaks = seq(1, 51, by = 5), limits = c(1, 51))
+
  scale_fill_viridis_d(option = "D") +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(
    title = "Relative Proportions of Pathogen Strains Over Time",
    x = "Weeks of Life",
```

```
    y = "Proportion of Pathogens (stacked to 100%)",
    fill = "Pathogen"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

Code for Appendix B Table 1

```
binary_pathogen_data_per_calf <- binary_pathogen_data %>%
  group_by(calf_id) %>%

  summarise(across(all_of(pathogen_cols), ~ as.integer(any(. > 0))))

# Now add in aetiological cause
cause <- final_miseq_data_clean %>%
  select(calf_id, definitive_aetiological_cause) %>%
  distinct()

binary_pathogen_data_per_calf_annotated <-
binary_pathogen_data_per_calf %>%
  left_join(cause, by = "calf_id")

# Count the number of pathogens each calf is infected with
infection_counts <- binary_pathogen_data_per_calf %>%
  mutate(n_infections = rowSums(across(all_of(pathogen_cols)))) %>%
  mutate(infection_category = case_when(
    n_infections == 0 ~ "No infections",
    n_infections == 1 ~ "Single infection",
    n_infections == 2 ~ "Dual infections",
    n_infections >= 3 ~ "Triple or more infections"
  ))

# Summarise into your required table
infection_summary <- infection_counts %>%
  group_by(infection_category) %>%
  summarise(
    n_calves = n(),
    percent_of_total = round((n() / nrow(.)) * 100, 1),
    .groups = "drop"
  ) %>%
```

```
  arrange(factor(infection_category, levels = c("No infections",
"Single infection", "Dual infections", "Triple or more infections")))
```

Code for Appendix B Figure 1

```
binary_pathogen_data_per_calf <- binary_pathogen_data_per_calf %>%
  rowwise() %>%
  mutate(n_unique_pathogens = sum(c_across(all_of(pathogen_cols))))
%>%
  ungroup()

binary_pathogen_data_per_calf <- binary_pathogen_data_per_calf %>%
  left_join(select(final_miseq_data_clean, calf_id,
definitive_aetiological_cause) %>% distinct(), by = "calf_id")

# Sum read counts across all visits per calf
calf_total_load <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(
    total_pathogen_load = sum(across(all_of(pathogen_cols)), na.rm =
TRUE),
    .groups = "drop"
  )

binary_pathogen_data_per_calf <- binary_pathogen_data_per_calf %>%
  left_join(calf_total_load, by = "calf_id")


ggplot(binary_pathogen_data_per_calf, aes(x = n_unique_pathogens, y =
total_pathogen_load)) +
  geom_boxplot()

ggplot(binary_pathogen_data_per_calf, aes(x =
as.factor(n_unique_pathogens), y = total_pathogen_load, fill =
as.factor(n_unique_pathogens))) +
  geom_violin(trim = TRUE, alpha = 0.7, color = NA) +
  geom_boxplot(width = 0.1, outlier.shape = NA, color = "white")+
  scale_fill_viridis_d(option = "D", name = "No. of Pathogens") +
  scale_y_continuous(
    trans = "log10",
    labels = scales::scientific,  # <--- Use scientific notation
    breaks = c(1e3, 1e4, 1e5, 1e6, 1e7, 1e8)
  ) +
```

```
  labs(
    title = "Total Pathogen Load by Number of Unique Pathogens",
    x = "Number of Unique Pathogens",
    y = "Total Read Count (log10 scale)"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.x = element_text(size = 11),
    axis.text.y = element_text(size = 10)
  )

cor.test(binary_pathogen_data_per_calf$total_pathogen_load,
binary_pathogen_data_per_calf$n_unique_pathogens, method =
"spearman")

#CI
spearman_rho <- function(data, indices) {
  d <- data[indices, ]
  cor(d$total_pathogen_load, d$n_unique_pathogens, method =
"spearman")
}

# Prepare data
df <- binary_pathogen_data_per_calf[, c("total_pathogen_load",
"n_unique_pathogens")]

# Bootstrap with 2000 resamples
set.seed(123)
boot_res <- boot(df, spearman_rho, R = 2000)

# Get 95% CI
boot.ci(boot_res, type = "perc")
```

Code for Appendix B Table 2

```
# Create a label for each unique infection combination per calf
co_infection_summary <- binary_pathogen_data_per_calf_annotated %>%
  mutate(combo = apply(select(., all_of(pathogen_cols)), 1,
function(x) {
    if (sum(x) == 0) {
      return("None")
```

```
    } else {
      return(paste(sort(names(x)[x == 1]), collapse = " + "))
    }
  })) %>%
  group_by(combo) %>%
  summarise(
    n_calves = n(),
    deaths = sum(definitive_aetiological_cause == "Dead"),
    survival = sum(definitive_aetiological_cause == "Alive"),
    mortality_rate = round(deaths / n_calves * 100, 1),
    .groups = "drop"
  ) %>%
  mutate(percent_of_total = round(n_calves / sum(n_calves) * 100, 1))

# Arrange in table
co_infection_summary %>%
  arrange(desc(n_calves))

# Store the total number of calves before grouping
total_calves <- sum(co_infection_summary$n_calves)

# Now collapse and calculate % using original total
co_infection_summary_collapsed <- co_infection_summary %>%
  mutate(combo_grouped = ifelse(n_calves < 6, "Other (rare)", combo))
%>%
  group_by(combo_grouped) %>%
  summarise(
    n_calves = sum(n_calves),
    deaths = sum(deaths),
    mortality_rate = round(deaths / n_calves * 100, 1),
    percent_of_total = round(n_calves / total_calves * 100, 1),  #
use full total here!
    .groups = "drop"
  ) %>%
  arrange(desc(n_calves))

# Create a combination label for each calf
co_infection_table <- binary_pathogen_data_per_calf_annotated %>%
  mutate(
    combo = apply(select(., all_of(pathogen_cols)), 1, function(x) {
      if (sum(x) == 0) {
        return("None")
      } else {
        return(paste(sort(names(x)[x == 1]), collapse = " + "))
      }
```

```
    })
  )

# Summarise how many calves have each pattern and their outcomes
co_infection_summary <- co_infection_table %>%
  group_by(combo) %>%
  summarise(
    n_calves = n(),
    deaths = sum(definitive_aetiological_cause == "Dead"),
    survival = sum(definitive_aetiological_cause == "Alive"),
    mortality_rate = round(deaths / n_calves * 100, 1),
    percent_of_total = round(n_calves / sum(n_calves) * 100, 1),
    .groups = "drop"
  ) %>%
  arrange(desc(n_calves))
```

Code for Appendix B Figure 2 & 3

```
# By sample week
upset(
  binary_pathogen_data,
  intersect = theileria_cols,
  min_size = 50,
  width_ratio = 0.3,
  base_annotations = list(
    'Intersection size' = intersection_size(
      mapping = aes(fill = SampleWeek)
    )
  )
) +
  scale_fill_viridis_d(option = "D", name = "Sample Week") +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),   # Remove bottom axis text
    axis.ticks.x = element_blank(),  # Remove tick marks
    plot.margin = margin(10, 10, 10, 10)  # Optional: add space
  )

# By survival outcome
binary_pathogen_data_per_calf <- binary_pathogen_data %>%
  group_by(calf_id) %>%
  summarise(across(all_of(theileria_cols), ~ as.integer(any(. > 0))))
```

```
# Now add in aetiological cause
cause <- binary_pathogen_data %>%
  select(calf_id, definitive_aetiological_cause) %>%
  distinct()

binary_pathogen_data_per_calf_annotated <-
binary_pathogen_data_per_calf %>%
  left_join(cause, by = "calf_id")

upset(
  binary_pathogen_data_per_calf_annotated,
  intersect = theileria_cols,
  min_size = 5,
  width_ratio = 0.4,
  set_sizes = upset_set_size(),
  base_annotations = list(
    'Intersection size' = intersection_size(
      mapping = aes(fill = definitive_aetiological_cause)
    ) +
      scale_fill_viridis_d(name = "Survival outcome", option = "D")
  )
)

# By sample week
upset(
  binary_pathogen_data,
  intersect = anaplasma_cols,
  min_size = 50,
  width_ratio = 0.3,
  base_annotations = list(
    'Intersection size' = intersection_size(
      mapping = aes(fill = SampleWeek)
    )
  )
) +
  scale_fill_viridis_d(option = "D", name = "Sample Week") +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),   # Remove bottom axis text
    axis.ticks.x = element_blank(),  # Remove tick marks
    plot.margin = margin(10, 10, 10, 10)  # Optional: add space
  )

# By survival outcome
binary_pathogen_data_per_calf <- binary_pathogen_data %>%
```

```
  group_by(calf_id) %>%
  summarise(across(all_of(anaplasma_cols), ~ as.integer(any(. > 0))))

# Now add in aetiological cause
cause <- binary_pathogen_data %>%
  select(calf_id, definitive_aetiological_cause) %>%
  distinct()

binary_pathogen_data_per_calf_annotated <-
binary_pathogen_data_per_calf %>%
  left_join(cause, by = "calf_id")

upset(
  binary_pathogen_data_per_calf_annotated,
  intersect = anaplasma_cols,
  min_size = 5,
  width_ratio = 0.4,
  set_sizes = upset_set_size(),
  base_annotations = list(
    'Intersection size' = intersection_size(
      mapping = aes(fill = definitive_aetiological_cause)
    ) +
      scale_fill_viridis_d(name = "Survival outcome", option = "D")
  )
)
```

Code for Appendix B Table 3

```
binary_pathogen_data_per_calf <- binary_pathogen_data %>%
  group_by(calf_id) %>%

  summarise(across(all_of(pathogen_cols), ~ as.integer(any(. > 0))))

# Now add in aetiological cause
cause <- final_miseq_data_clean %>%
  select(calf_id, definitive_aetiological_cause) %>%
  distinct()

binary_pathogen_data_per_calf_annotated <-
binary_pathogen_data_per_calf %>%
  left_join(cause, by = "calf_id")

# Create simplified classification
```

```
tparva_status <- binary_pathogen_data_per_calf_annotated %>%
  mutate(
    tparva = theileria_parva_l02366_tb,
    tmutt = theileria_mutans_af078815_tb,
    co_infection = rowSums(across(all_of(pathogen_cols))) > 1,
    group = case_when(
      tparva == 1 & co_infection == FALSE ~ "T. parva only",
      tparva == 1 & co_infection == TRUE ~ "T. parva + co-infection",
      tparva == 0 ~ "No T. parva"
    )
  ) %>%
  group_by(group) %>%
  summarise(
    n_calves = n(),
    deaths = sum(definitive_aetiological_cause == "Dead"),
    mortality_rate = round(deaths / n_calves * 100, 1),
    .groups = "drop"
  )

# Create contingency table
table_tparva <- binary_pathogen_data_per_calf_annotated %>%
  mutate(
    tparva = theileria_parva_l02366_tb,
    co_infection = rowSums(across(all_of(pathogen_cols))) > 1,
    group = case_when(
      tparva == 1 & co_infection == FALSE ~ "T. parva only",
      tparva == 1 & co_infection == TRUE ~ "T. parva + co-infection",
      tparva == 0 ~ "No T. parva"
    ),
    died = ifelse(definitive_aetiological_cause == "Dead", 1, 0)
  ) %>%
  count(group, died) %>%
  pivot_wider(names_from = died, values_from = n, values_fill = 0)
%>%
  column_to_rownames("group")  # Needed for test

# Apply Fisher's Exact Test
fisher.test(as.matrix(table_tparva))
groups <- rownames(table_tparva)

pairwise_results <- lapply(combn(groups, 2, simplify = FALSE),
function(g) {
  sub_tab <- table_tparva[g, ]
  test <- fisher.test(as.matrix(sub_tab))
  list(
```

```
    comparison = paste(g, collapse = " vs "),
    p.value = test$p.value,
    odds.ratio = if (!is.null(test$estimate)) unname(test$estimate)
else NA,
    conf.int = if (!is.null(test$conf.int)) test$conf.int else c(NA,
NA)
  )
})

pairwise_results
```

Code for Appendix B Table 4

```
# Summary table
table <- tbl_cross(data = windowed_data, row = infection_order, col =
event, percent = "row")
table
```

Code for Appendix B Figure 4

```
# Identify earliest infection for each pathogens
earliest_haemo <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarize(
    earliest_mutans =
suppressWarnings(min(sample_week[theileria_mutans_af078815_tb > 0],
na.rm = TRUE)),
    earliest_velifera =
suppressWarnings(min(sample_week[theileria_velifera_af097993_tb > 0],
na.rm = TRUE)),
    earliest_parva =
suppressWarnings(min(sample_week[theileria_parva_l02366_tb > 0],
na.rm = TRUE)),
    .groups = "drop"
  ) %>%
  mutate(across(starts_with("earliest_"), ~ ifelse(is.infinite(.),
51, .)))  # Convert Inf to NA

earliest_haemo <- earliest_haemo %>%
  pivot_longer(cols = 2:4, names_to = "pathogen_strain", values_to =
"initial_infection_week")
```

```
# Assign  event status (1 = infected with pathogen, 0 = not)
earliest_haemo$event <- ifelse(earliest_haemo$initial_infection_week
== "51", 0, 1)

# Create the survival object
surv_obj <- Surv(time = earliest_haemo$initial_infection_week, event
= earliest_haemo$event)

# Relabel strain names
earliest_haemo <- earliest_haemo %>%
  mutate(pathogen_strain = recode(pathogen_strain,
                                  "earliest_parva" = "T. parva",
                                  "earliest_mutans" = "T. mutans",
                                  "earliest_velifera" = "T. velifera"
  ))

# Fit Kaplan-Meier curves grouped by parasite
fit <- survfit(surv_obj ~ pathogen_strain, data = earliest_haemo)

viridis_palette <- viridis(3, option = "D")

# Plot the survival curves
ggsurvplot(
  fit,
  data = earliest_haemo,
  risk.table = TRUE,
  pval = TRUE,
  conf.int = TRUE,
  pval.method = TRUE,
  pval.size = 5,
  pval.coord = c(12, 0.30),
  censor = TRUE,
  palette = viridis_palette,
  ggtheme = theme_minimal(base_size = 14),
  xlab = "Time to First Infection (weeks)",
  ylab = "Proportion Not Yet Infected",
  legend.title = "Parasite Group",
  legend.labs = c("T. mutans", "T. parva", "T. velifera")  # remove
variable name prefix
)
```

Code for Appendix B Figure 5

```
# Filter to only calves that died
died_data <- windowed_data_detailed %>%
  filter(event == 1)

# Check how many in each group
table(died_data$infection_order)

# Compare time to event (death) between groups
wilcox.test(time_to_event ~ infection_order, data = died_data)

ggplot(died_data, aes(x = infection_order, y = time_to_event, fill =
infection_order)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, outlier.shape = NA, color = "black") +
  labs(
    title = "Time to Death by Infection Order (Only Calves that
Died)",
    x = "Infection Order",
    y = "Time to Death (weeks)"
  ) +
  scale_fill_viridis_d(option = "D") +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```

Code for Appendix C Table 1

```
# Create a contingency table for genotype and survival status (died)
contingency_table <- table(Genomics_clean$genotype,
Genomics_clean$died)

table <- tbl_cross(data = Genomics_clean, row = genotype, col = died,
percent = "row")
table

# Perform Fisher's Exact Test
fisher_test <- fisher.test(contingency_table)
print(fisher_test)
```

Code for Appendix C Table 2

```
# Identify if each calf was ever infected with T. parva
tparva_infection <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
```

```
  summarise(tparva_positive = any(theileria_parva_l02366_tb > 0,
na.rm = TRUE))

# Merge with genotype
tparva_infection <- tparva_infection %>%
  left_join(Genomics_clean %>% select(calf_id, genotype), by =
"calf_id")

infection_rate_tbl <- tparva_infection %>%
  filter(genotype %in% c("TT", "CT", "CC")) %>%
  group_by(genotype) %>%
  summarise(
    n_calves = n(),
    n_infected = sum(tparva_positive, na.rm = TRUE),
    perc_infected = round(n_infected / n_calves * 100, 1)
  )

print(infection_rate_tbl)
```

Code for Appendix C Figure 1

```
# Sum total T. parva reads per calf
tparva_load <- final_miseq_data_clean %>%
  group_by(calf_id) %>%
  summarise(tparva_load = sum(theileria_parva_l02366_tb, na.rm =
TRUE)) %>%
  left_join(Genomics_clean %>% select(calf_id, genotype), by =
"calf_id") %>%
  filter(genotype %in% c("TT", "CT", "CC"))

kruskal.test(tparva_load ~ genotype, data = tparva_load)
tt_cc_only <- tparva_load %>% filter(genotype %in% c("TT", "CC"))
wilcox.test(tparva_load ~ genotype, data = tt_cc_only)

ggplot(tparva_load, aes(x = genotype, y = tparva_load, fill =
genotype)) +
  geom_violin(trim = FALSE, alpha = 0.8) +
  geom_boxplot(width = 0.1, outlier.shape = NA, color = "white") +
  scale_y_log10(
    labels = label_number(scale_cut = cut_short_scale()),
    breaks = trans_breaks("log10", function(x) 10^x)
  ) +
  scale_fill_viridis_d(option = "D") +
  labs(
```

```
  title = "T. parva Pathogen Load by Genotype",
  x = "Genotype",
  y = "Total T. parva Read Count (log scale)"
) +
theme_minimal(base_size = 14)
```