Mining the Genetic Diversity of the Pathogenic Protozoan, *Neospora caninum*

By

Larissa Calarco

January
2020

School of Life Sciences
Faculty of Science
University of Technology Sydney

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) at the University of Technology Sydney (UTS)

Certificate of Original Authorship

I, Larissa Calarco, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Life Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Date: January 2020

Dedication

This thesis is dedicated to my wonderful parents, Dominic and Kathy Calarco. I would not be who or where I am today without their love, sacrifices, generosity, and constant support, not only throughout my studies, but everything leading up to this point. I am eternally grateful for the ample opportunities they have strived to provide me with, always ensuring that I have the means to reach my full potential.

Acknowledgements

I would like to extend my sincerest thanks and gratitude to my supervisors, Professor John Ellis and Dr Joel Barratt, who I am indebted to. Completion of this thesis and research degree would not have been impossible without their constant and unfaltering support, succour, encouragement, patience, and guidance. It has truly been an honour and privilege being part of the Ellis research group for the last four years, and being supervised by such eminent, renowned professionals in their field has been an extremely rewarding and humbling experience.

I would like to thank Alexa Kaufer, Rory Gough, and the other members of the Ellis research group for their support, comradery, solidarity, and assistance. It's been a pleasure working alongside all of you.

I would also like to thank the University of Technology Sydney for the generous PhD scholarship, which has supported me throughout the course of my degree.

Last by not least, I would like to thank the ongoing support, love, and encouragement of my parents, grandparents, immediate and extended families, and friends.

Table of Contents

Certificate of Original Authorship i
Dedication ii
Acknowledgements iii
Table of Contents iv
Publications arising from thesis v
Conference presentations arising from thesis vi
List of Figures vii
List of Tables ix
Supplementary Information x
Thesis Abstract xi
Exegesis xiii
Chapter 1 1
Chapter 2 20
Chapter 3 36
Chapter 4 52
Chapter 5 62

Publications arising from thesis

Calarco, L., Barratt, J., Ellis, J., 2018. *Genome wide identification of mutational hotspots in the apicomplexan parasite Neospora caninum and the implications for virulence*. **Genome Biol Evol**, 10 (9), 2417-2431.

Calarco, L. and Ellis, J., 2019. *Annotating the 'hypothetical' in hypothetical proteins: in-silico analysis of uncharacterised proteins for the Apicomplexan parasite, Neospora caninum*. **Vet Parasitology**, 265, 29-37.

Calarco, L., Barratt, J., Ellis, J., 2020. *Detecting sequence variants in clinically important protozoan parasites*. **Int J Parasitol,** 50 (1), 1-18.

Calarco, L. and Ellis, J., 2020. *Contribution of introns to the species diversity associated with the apicomplexan parasite, Neospora caninum*. **Parasitol Res**, 1-15.

Conference presentations arising from thesis

Calarco, L., Bush, S., Barratt, J. & Ellis, J. *Validation of a variant analysis pipeline for the identification of vaccine candidates for eukaryotic pathogenic organisms*. International Congress for Tropical Medicine and Malaria; Brisbane, Australia (2016).

Calarco, L., Barratt, J. & Ellis, J. *Validation of a variant analysis pipeline for the identification of vaccine candidates for eukaryotic pathogenic organisms*. Australian Society for Parasitology Annual Conference; Leura, Australia (2017).

Calarco, L., Barratt, J. & Ellis, J. *Genome wide identification of mutational hotspots in the apicomplexan parasite Neospora caninum, and the implications for virulence*. 14th International Congress of Parasitology; Daegu, Korea (2018).

Calarco, L., Barratt, J. & Ellis, J. *Genome wide identification of mutational hotspots in the apicomplexan parasite Neospora caninum, and the implications for virulence.* Australian Society for Parasitology Annual Conference; St Kilda, Australia (2018).

Calarco, L. & Ellis, J. *Contribution of introns to the species diversity associated with the apicomplexan parasite, Neospora caninum*. Australian Society for Parasitology Annual Conference; Adelaide, Australia (2019).

List of Figures

Chapter 1……..1

Figure 1. Summary of a typical variant detection workflow for next generation and third generation sequencing data……..5

Figure 2. Example of false-positive and true variants identified by a variant analysis and visualised in the Integrative Genomics Viewer……..9

Figure 3. Visual summary of the types of sequence variants and large structural variants……13

Chapter 2……20

Figure 1. A summary of the complete variant detection workflow used in this study……25

Figure 2. Circos plot representing the SNP data in the context of the *N. caninum* genome…….....27

Figure 3. Circos plot presenting the location of non-synonymous SNPs identified in the *N. caninum* genome…….29

Figure 4. Plots displaying the features of chromosome VI (A) and XI (B)…….30

Figure 5. Unrooted neighbour-joining trees showing the population structure within the *Neospora* genus……31

Chapter 3……36

Figure 1. Various SNPs identified by VarScan and visualised in IGV, that were subsequently allocated to introns within *N. caninum* genes…….42

Figure 2. A Circos plot presenting the SNPs identified by VarScan, in the context of the *N. caninum* genome……43

Figure 3. Summary of the methodology and main findings pertaining to the comparative transcriptome and genome analyses……45

Figure 4. Schematic of the annotated *N. caninum* apicoplast sequence assembled with ABySS, based on genome data from NC-Liverpool…...46

Chapter 4……52

Figure 1. A summary of the workflow used to annotate uncharacterised proteins for *N. caninum*……...55

Figure 2. Classification of all 3283 uncharacterised proteins under investigation by Philius……..56

Figure 3. Pie of a pie chart presenting the predicted topology for uncharacterised proteins predicted to be adhesins……...56

Chapter 5……62

Figure 1. Variant analysis pipeline for detecting SNPs and indels, from raw high-throughput sequencing data generation through to functional analysis of identified variants……...74

Figure 2. Comparison of the gene catalogue and structure of the *N. caninum* and *T. gondii* apicoplast genomes……86

List of Tables

Chapter 1……..1

Table 1. Features of read alignment tools……..6

Table 2. Features of popular variant calling software…….....7

Table 2. List of recommended filtering strategies to obtain high confidence callsets, following *in silico* detection…….....8

Chapter 2……20

Table 1. Summary of VarScan variant calling using NC-Nowra RNA-Seq reads aligned to the de novo NC-Liverpool transcriptome reference……26

Table 2. A summary of the total number of variants selected for and confirmed through sequencing, based on targeting various loci…….26

Chapter 4……52

Table 1. Summary of the tools used in the annotation of uncharacterised *N. caninum* proteins……...54

Table 2. Annotations for the final 32 proteins, including their gene descriptions, InterProScan results, and gene ontologies…….....57

Supplementary Information

The thesis chapters 2-4 make reference to supplementary files, online resources, tables, and spreadsheets, containing additional supporting information related to respective methods and results. This information is contained within files accessible either at the URL specified at the end of published chapters, or on the Supplementary Information disk provided.

Thesis Abstract

Neospora caninum is a cyst-forming apicomplexan parasite, responsible for economic and reproductive losses to cattle industries worldwide, and represents a serious neurological disease in canines. Although discovered over three decades ago, progress towards treatment and control strategies against neosporosis, remains stagnant. Currently, common practices to combat the disease include passivity, or expensive culling of seropositive dams. However, vaccination represents a cost-effective and efficacious option, especially using live, attenuated isolates.

Members of the Apicomplexa consist of populations that vary enormously in their disease-causing potential, where *in vivo* experiments have demonstrated pathogenic variability between *N. caninum* isolates. The underlying question therefore, is what is the genetic basis of virulence within the species, and consequently, how can such information be exploited in vaccine development? Thus far, conventional techniques have been employed to study the intraspecies genetic diversity associated with *N. caninum*, generally involving PCR-based approaches targeting repetitive elements. However, a direct causal relationship between such diversity and important parasite phenotypes such as virulence, is yet to be established.

Alternatively, burgeoning next generation sequencing (NGS) technologies and *in-silico* tools have provided new opportunities to perform genome-wide scans in such organisms. Hence the objective of this body of work was to compare the genomes and transcriptomes of two distinct *N. caninum* isolates, using NGS data and bioinformatics workflows, to identify sequence variants in coding and non-coding DNA. Annotation of variable regions would reveal potential virulence markers distinguishing isolates of this species. Challenges accompanying such research include the lack of optimisation and

standardisation of NGS analysis tools for non-model organisms such as pathogenic Protozoa. This is compounded by the dubious accuracy of the *N. caninum* reference genome, as well as the disturbingly large number of proteins described as 'hypothetical' or 'uncharacterised'.

This body of research represents a thesis by compilation, consisting of four publications, and one chapter under review. Each chapter represents an independent study, which collectively address the research objective and current gaps in the literature. The results present polymorphic "hotspots" in concentrated windows of the *N. caninum* genome, where there is a correlation between hypervariable regions within protein-coding genes, and non-coding regions. Furthermore, an *in-silico* pipeline is developed to annotate uncharacterised proteins, subsequently identifying a subset of proteins potentially implicated in crucial parasite mechanisms, conducive to *N. caninum*'s success. It is trusted that this thesis contributes vital knowledge pertaining to *N. caninum* intraspecies diversity, aiding in the quest to develop a vaccine against neosporosis.

Exegesis

Neospora caninum is a protozoan parasite of economic and veterinary significance. Despite it being a close relative of the well-studied, ubiquitous, model apicomplexan, *Toxoplasma gondii*, the current understanding of genetic drivers of phenotypic diversity between isolates of *N. caninum*, leaves much to be desired. The main objective of this body of research, which represents a thesis by compilation, focused on elucidating unexplored sources of genetic diversity amongst isolates of *N. caninum*, and determining genotypic and phenotypic relationships of biologically significant variable loci. Each chapter of this thesis is an independent sub-study with relevant literature, methodologies, results, and discussions, where collectively they address the principal, multi-faceted research objective:

Chapter 1 presents a literature review of the current status of Next Generation Sequencing (NGS) data analysis, with a focus on the identification and exploitation of sequence variants. The ever increasing availability, throughput, and affordability of NGS technology, has afforded new ways to explore and compare the genomes of organisms, with applications in population genetics, disease, and evolutionary diversity. However, advancements in bioinformatics and *in-silico* workflows designed to analyse such data in a biologically significant context, are plagued with limitations and challenges, especially for non-model organisms such as pathogenic Protozoa. This review provides recommendations on implementing and optimising available workflows for parasitic Protozoa, and emphasises how such data can and has been exploited.

Chapter 2 focuses on elucidating the presence and impact of small sequence variants located within protein-coding genes of *N. caninum*. This research used RNA-Seq data from two distinct *N. caninum* isolates, for *in-silico* variant detection, the results of which were validated through laboratory analysis, and visualised in the context of the *N. caninum* genome. This work was also extended to additional *Neospora* isolates which differ by host,

reported pathogenicity and/or geographical distribution, to explore underlying population genetics distinguishing members of this genus.

Chapter 3 extends the results of the preceding chapter, by exploiting Whole Genome Sequencing (WGS) data to explore the contribution of variable non-coding DNA and introns to *N. caninum* parasite diversity. This research reports the intersection of SNP-dense coding and non-coding regions, and annotates genes within these distinct, prioritised genomic windows. The first assembled and annotated apicoplast genome for *N. caninum* is also presented, where this essential organelle is shown to be highly conserved within *N. caninum* and related Coccidia. Overall, this work presents additional sources contributing to the intraspecies genetic diversity in *N. caninum*, and highlights the importance of non-coding DNA when studying biologically significant phenotypes such as virulence.

Chapter 4 describes an *in-silico* workflow for the annotation and prioritisation of hypothetical or uncharacterised proteins for *N. caninum*. This work was prompted by the challenges associated with the previous research chapters, with respect to inferring the biological implications and functional impact of genomic and transcriptome variation identified between isolates. This workflow not only identified a subset of previously uncharacterised *N. caninum* proteins potentially contributing to crucial parasite mechanisms, but also has applicability for other non-model organisms that lack complete, adequate, or robust genome annotations.

Finally, **Chapter 5** contains a cohesive discussion reviewing the main findings of each research chapter independently and as they relate to one another, and contextualises the results against the backdrop of the current state of *N. caninum* research. It also suggests potential future avenues for research to combat, treat, and control neosporosis. In summary, this chapter highlights how this unified body of work contributes novel, timely, and vital research towards illuminating potential new targets for control options against neosporosis.

Chapter 1

Detecting sequence variants in clinically important protozoan parasites

Publication:

Calarco, L., Barratt, J., Ellis, J., 2020. *Detecting sequence variants in clinically important protozoan parasites*. **Int J Parasitol,** 50 (1), 1-18.

Certificate:

I certify that the following chapter is largely my own work, although the contributions of other authors are duly recognised. The contribution of other authors are detailed as follows:

- By providing suggestions on topics to be reviewed
- By suggesting references and works to be reviewed as part of this work
- By proof reading draft manuscripts
- By correcting spelling and grammatical errors in drafts
- By providing suggestions to improve writing style and language
- By providing suggestions to improve layout and structure

Otherwise, the core composition of this work is credited to me.

I hereby certify that the above statements are true and correct:

Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.
Larissa Calarco	Dr Joel Barratt	Prof John Ellis
PhD Candidate	Co-Author	Co-Author

Date: January 2020

International Journal for Parasitology 50 (2020) 1–18

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Review Article

Detecting sequence variants in clinically important protozoan parasites

Larissa Calarco *, Joel Barratt, John Ellis

School of Life Sciences, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 19 June 2019
Received in revised form 29 September 2019
Accepted 1 October 2019
Available online 16 December 2019

Keywords:
Variant analysis
Population genetics
Evolutionary selection
NGS data
Non-model organism
Structural variants
SNPs

ABSTRACT

Second and third generation sequencing methods are crucial for population genetic studies, and variant detection is a popular approach for exploiting this sequence data. While mini- and microsatellites are historically useful markers for studying important Protozoa such as *Toxoplasma* and *Plasmodium* spp., detecting non-repetitive variants such as those found in genes can be fundamental to investigating a pathogen's biology. These variants, namely single nucleotide polymorphisms and insertions and deletions, can help elucidate the genetic basis of an organism's pathogenicity, identify selective pressures, and resolve phylogenetic relationships. They also have the added benefit of possessing a comparatively low mutation rate, which contributes to their stability. However, there is a plethora of variant analysis tools with nuanced pipelines and conflicting recommendations for best practise, which can be confounding. This lack of standardisation means that variant analysis requires careful parameter optimisation, an understanding of its limitations, and the availability of high quality data. This review explores the value of variant detection when applied to non-model organisms such as clinically important protozoan pathogens. The limitations of current methods are discussed, including special considerations that require the end-users' attention to ensure that the results generated are reproducible, and the biological conclusions drawn are valid.

© 2019 Australian Society for Parasitology. Published by Elsevier Ltd. All rights reserved.

[Production Note: This paper is not included in this digital copy due to copyright restrictions.]

View/Download from: Publisher's site

* Corresponding author at: School of Life Sciences, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia. Fax: +61 2 9514 8206.
E-mail address: Larissa.M.Calarco@student.uts.edu.au (L. Calarco).

https://doi.org/10.1016/j.ijpara.2019.10.004
0020-7519/© 2019 Australian Society for Parasitology. Published by Elsevier Ltd. All rights reserved.

Chapter 2

Genome wide identification of mutational hotspots in the apicomplexan parasite *Neospora caninum*, and the implications for virulence

---

Publication:

Calarco, L., Barratt, J., Ellis, J. 2018. *Genome wide identification of mutational hotspots in the apicomplexan parasite Neospora caninum and the implications for virulence*. **Genome Biol Evol**, 10 (9), 2417-2431.

Certificate:

I certify that the following chapter is largely my own work, although the contributions of other authors are duly recognised. The contribution of other authors are detailed as follows:

- By providing suggestions on topics to be reviewed
- By suggesting references and works to be reviewed as part of this work
- By proof reading draft manuscripts
- By correcting spelling and grammatical errors in drafts
- By providing suggestions to improve writing style and language
- By providing suggestions to improve layout and structure

Otherwise, the core composition of this work is credited to me.

I hereby certify that the above statements are true and correct:

Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.
Larissa Calarco	Dr Joel Barratt	Prof John Ellis
PhD Candidate	Co-Author	Co-Author

Date: January 2020

Genome Wide Identification of Mutational Hotspots in the Apicomplexan Parasite *Neospora caninum* and the Implications for Virulence

Larissa Calarco*, Joel Barratt, and John Ellis

School of Life Sciences, University of Technology Sydney, New South Wales, Australia

***Corresponding author:** E-mail: larissa.m.calarco@student.uts.edu.au.

Accepted: August 23, 2018

Data deposition: This project has been deposited at GenBank and SRA, with accessions MH719238-MH719327 and SRP156994 respectively.

Abstract

Neospora caninum is an apicomplexan parasite responsible for neosporosis, a disease causing hind limb paralysis in dogs and abortion in cattle, resulting in substantial economic losses to beef and dairy industries. Marked differences in pathogenicity exist between *N. caninum* strains suggesting that intrinsic genetic differences exist between them. These differences likely exist in genes expressed during the tachyzoite lifecycle stage which is responsible for the pathogenesis of neosporosis. An improved understanding of these genetic differences is essential to understanding *N. caninum* virulence, though such knowledge is scarce. Using a variant detection workflow we compared the tachyzoite transcriptomes of two *N. caninum* strains with different virulence properties: NC-Liverpool (virulent) and NC-Nowra (avirulent). This workflow identified 3130 SNPs and 6123 indels between the strains, and nine markers capturing 30 variants were Sanger sequenced for both strains. Sequencing of these loci was extended to an additional eight strains and subsequent phylogenetic analysis supported a genetic population structure comprised of two major clades with no geographical segregation. Sequence polymorphisms within coding regions of tachyzoite-associated genes were concentrated on chromosomes XI and XII, with 19 distinct tachyzoite-associated SNP hotspot regions identified within coding regions of the *N. caninum* nuclear genome. The variants were predominantly located in loci associated with protein binding, protein–protein interactions, transcription, and translation. Furthermore, 468 nonsynonymous SNPs identified within protein-coding genes were associated with protein kinase activity, protein binding, protein phosphorylation, and proteolysis. This work may implicate these processes and the specific proteins involved as novel effectors of *N. caninum* tachyzoite virulence.

Key words: variant analysis, SNP hotspot, transcriptomics, nonsynonymous mutations, population structure.

Introduction

Neospora caninum is a cyst forming coccidian of the phylum Apicomplexa first described as the cause of a potentially fatal neurological disease of dogs (Dubey et al. 1988a). However, its economic importance is primarily due to its role as the etiological agent of bovine neosporosis, a reproductive disease characterized by abortion and stillbirths in cows that is recognized as the leading global cause of bovine reproductive failure (Dubey and Lindsay 1996; Dubey 1999; Reichel and Ellis 2002; Dubey and Dubey 2003; Dubey et al. 2006; Reichel et al. 2007). Bovine infections with *N. caninum* have been reported in the Americas, Europe, Australia, and New Zealand, causing losses within the range of US $1.1 million in New Zealand, to an average total of US $546.3 million in the USA (Reichel et al. 2013). The combined annual losses due to *N. caninum* for the Australian and New Zealand dairy and beef industries are estimated to be greater than AU $110 million annually (Miller et al. 2002; Reichel and Ellis 2002).

Neospora caninum is a diverse species and several strains have been characterized revealing notable genotypic and phenotypic differences (Al-Qassab et al. 2010b). For example, the highly virulent NC-Liverpool strain causes foetal death in cattle (Atkinson et al. 1999), whereas the NC-Nowra strain has been evaluated for use as a live attenuated vaccine

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

against bovine neosporosis, based on its low virulence in mouse models (Miller et al. 2002; Williams et al. 2007; Weber et al. 2013). NC-Liverpool infection in mice causes severe neosporosis, characterized by encephalitis, hind limb paralysis, and severe weight loss, whereas a Swedish bovine isolate, NC-SweB1, induces similar but significantly milder symptoms in a smaller number of infected mice (Atkinson et al. 1999). The NC1 strain of *N. caninum* is known to induce severe clinical manifestations including fetal death in cattle as well as polyradiculoneuritis and granulomatous polymyositis in infected dogs (Dubey et al. 1988a, 1988b, 1992; Innes et al. 2001; Maley et al. 2003). However, while there are studies that report marked differences in pathogenicity between *N. caninum* strains in mouse models, there are limited results published comparing the behavior of various strains in cattle. A study focusing on correlating fetal loss with *N. caninum* infection reported fetal death in pregnant heifers following inoculation of the BPA1 isolate at 118 days gestation (Barr et al. 1994). Furthermore, an absence of fetal death was reported in pregnant heifers inoculated with NC-Spain 1 H, an isolate of low virulence, whereas fetal death occurred in heifers inoculated with the control strain NC1 (Rojo-Montejo et al. 2009). Differences have also been demonstrated between virulent isolates NC-Spain 7 and NC1 in cattle, with respect to the timing of fetal death and immunological response, where NC-Spain 7 resulted in higher fetal mortality rates and an earlier and higher anti-*N. caninum* IgG response (Caspe et al. 2012).

These phenotypic differences reflect a genetically diverse species. Analysis of mini- and microsatellite repeats for over 100 *N. caninum* strains has revealed extensive genetic diversity (Regidor-Cerrillo et al. 2006, 2013; Basso et al. 2009; Al-Qassab et al. 2010a). A typing method based on randomly amplified polymorphic DNA (RAPD) resolved several *N. caninum* isolates into six genotypes (Schock et al. 2001). Additionally, Regidor-Cerrillo et al. (2006) performed multilocus microsatellite analysis of nine cultured *N. caninum* isolates with varying host ranges and geographical locations, which revealed distinct genetic profiles for the 12 microsatellite markers investigated. Similarly, a multiplex PCR targeting three microsatellites and three minisatellites (Tand-3, Tand-12, Tand-13, Cont-6, Cont-14, and Cont-16), was developed by typing 25 cultured *N. caninum* isolates which identified 11 genotypes (Al-Qassab et al. 2010a). Although these methods reflect the diversity of *N. caninum* as a species, they are based on repetitive sequences that are generally noncoding and their impact on parasite phenotype is unknown.

The Apicomplexa have evolved several unique features that aid them in their intracellular parasitic lifestyle. These include molecules that facilitate motility, host cell adhesion, and invasion. Apicomplexan parasites manipulate host cells through secretion of effector proteins produced by specialized secretory organelles unique to this phylum; micronemes, rhoptries, and dense granules (English et al. 2015). Micronemal (*MIC*) proteins are released upon contact with host cells and facilitate adhesion (Cerede et al. 2005), where for example *MIC2* plays a role in host-cell attachment, motility, and invasion in *T. gondii* (Lovett et al. 2000; Huynh and Carruthers 2006), and *MIC1* and *MIC3* are soluble adhesins Naguleswaran et al. 2001; Keller et al. 2002; Cerede et al. 2005). Rhoptry family proteins are then secreted into the host cell cytosol facilitating formation of the tight junction between the invading parasite and target host cell, culminating in the formation of the parasitophorous vacuole (Talevich and Kannan 2013). Shortly after host cell invasion, the dense granules release *GRA* proteins that may be involved in nutrient acquisition (Nam 2009; Leineweber et al. 2017). Studies of the closely related apicomplexan parasite *Toxoplasma gondii* have identified a range of virulence factors that exist as orthologues in *N. caninum*, including dense granule protein *GRA9* (Leineweber et al. 2017), *ROP5* (Reese et al. 2011; Ma et al. 2017b), and *ROP16* and *ROP18* (Saeij et al. 2006; Taylor et al. 2006; Lei et al. 2014; Ma et al. 2017a).

Although current typing approaches for *N. caninum* have confirmed genetic variation in repetitive elements, there is a lack of knowledge on polymorphisms occurring in the coding regions of its genome. Sequence polymorphisms within many notable virulence factors have been described in *T. gondii*. For example, the identification of sequence polymorphisms within *GRA6* and *GRA7* of *T. gondii* led to the development of serotyping technology that is now commonly used for genotyping strains within this species (Kong et al. 2003; Sousa et al. 2009). Similarly, differences in virulence properties reported between *N. caninum* strains might imply that genetic diversity exists within, upstream, or downstream of genes that are transcriptionally active in tachyzoites which are the life cycle stage responsible for the pathogenesis of neosporosis.

The present study employed a variant detection workflow to compare the transcriptomes of two *N. caninum* strains with markedly different virulence properties: NC-Liverpool (virulent) and NC-Nowra (avirulent in mice). Phylogenetic analysis of sequenced polymorphic markers identified in silico, revealed a population structure consisting of two major clades showing no obvious geographical segregation. Tachyzoite-associated polymorphisms were associated with kinase activity, ATP binding, protein–protein interactions, and proteolysis, implicating several proteins involved in these processes as potentially novel determinants of *N. caninum* virulence.

Materials and Methods

Parasite Culture for Nucleic Acid Extraction and Sequencing

Neospora caninum strains (supplementary file S1, table S1, Supplementary Material online) were grown in vitro using Vero cells as the host cell line, at 37°C in RPMI media supplemented with 10% heat inactivated horse serum. Total RNA was extracted from the tachyzoites using TriSure reagent (Bioline) and treated with RNAase-free DNAase (Sigma).

RNA-seq was performed on three biological replicates of mRNA, each extracted from difference passages of both NC-Liverpool and NC-Nowra only. For each strain, two technical replicates (RNA-seq libraries) were prepared, constituting six libraries in total. The sequencing reads were generated using Illumina HiSeq2000, 100 base paired-end sequencing.

For laboratory confirmation of the SNPs, genomic DNA was also extracted using the solvent extraction technique, from cultured tachyzoites of *N. caninum* and *Neospora hughesi* strain NE1 (imported from ATCC). Briefly, cells were pelleted and then extracted three times with equal volumes of phenol and chloroform and then once more with chloroform only, with thorough vortexing and centrifuging at 13,000 × g for 1 min between extraction steps. The DNA was precipitated from the final aqueous phase by isopropanol and resuspended in 100 μl of ddH2O. The DNA extracts were stored at −20 °C until required.

Read Quality Control and Mapping

Illumina reads were trimmed for quality and length using the Filter FASTQ tool (Blankenberg et al. 2010a), available on the Galaxy Platform (Blankenberg et al. 2010b) through the Garvan Institute for Medical Research (http://galaxyproject.org/; Last accessed June 2015). Illumina reads <15 base pairs long, and with per base quality scores <20, were discarded using the Filter FASTQ tool (Bao et al. 2014; Pabinger et al. 2014; Broad Institute 2015). A Perl script (supplementary file S1, Supplementary Material online) was used to ensure paired read information was preserved, resulting in two paired read files, and an unpaired (singlet) read file. These processed reads were next mapped to the *N. caninum* reference genome available from ToxoDB (NC-Liverpool genome, version 28, http://www.toxodb.org/toxo/; Last Accessed April 2018) using TopHat version 2.1.1 (Bao et al. 2014; Pabinger et al. 2014; Broad Institute 2015). Read mapping was optimized by adjusting alignment parameters to increase the overall read alignment rate, as detailed in supplementary file S1, Supplementary Material online.

De Novo Transcriptome Assembly

An in-house reference transcriptome was created for NC-Liverpool by performing a de novo transcriptome assembly. The TopHat alignment tool was first used to map the *N. caninum* reads to the Vero genome, resulting in unmapped BAM files containing *N. caninum* reads that were sorted and converted into fastq files using scripts provided in supplementary file S1, Supplementary Material online. The resulting fastq files were assembled using Trinity (version 2.5.1) (Grabherr et al. 2011). Removal of redundant contigs was performed using CD-HIT-EST which sorts comparable nucleotide sequences based on a user-defined similarity threshold, and reports the longest sequence in each cluster as the representative contig (version 4.6.6) (Li and Godzik 2006). This step was included to ensure the same variants were not identified and duplicated in the final callset, within redundant contigs generated by Trinity from the same or very similar sequence reads. The scripts available in the Trinity package and the TransRate software package (version 1.0.3) (Smith-Unna et al. 2016) were employed to assess the quality of both the original and new transcriptome assembly following CD-HIT-EST analysis, using the parameters and thresholds provided in supplementary file S1, Supplementary Material online, where a similarity threshold of 0.8–0.85 was used ($n = 5$). A summary of assembly metrics assessed is contained within supplementary file S1, table S2, Supplementary Material online, and the NC-Liverpool transcriptome can be found in supplementary file S2, Supplementary Material online, in FASTA format. The NC-Liverpool transcriptome generated in-house was compared with published NC-Liverpool reference transcripts from ToxoDB by mapping RNA-seq reads generated in-house from NC-Liverpool. Variant calling was performed from the resulting BAM (binary alignment/map) files.

Variant Calling

SAMtools (Li et al. 2009) was used to sort and index the "mapped" BAM files generated by TopHat, and to generate an mpileup (multi-sample pileup) output. This data was then imported into VarScan 2 (Koboldt et al. 2013) for variant calling using the recommended parameters. The identified variants were filtered using VarScan's accessory scripts, which remove variants that do not meet thresholds pertaining to strand bias, sequence and variant coverage thresholds, mismatch quality sum, and read position bias. A detailed description of this workflow can be found in supplementary file S1, Supplementary Material online. Variants were visualized using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011; Thorvaldsdottir et al. 2013). Briefly, the *N. caninum* reference genome FASTA file was uploaded to IGV (version 2.3.67), along with the sorted BAM files for each sample. Hundreds of variants were randomly selected for viewing from the SNP and indel lists produced by VarScan, and a set of high confidence variants were selected from amongst these for laboratory validation.

Variant Annotation

The de novo transcriptome assembly generated for NC-Liverpool was queried against the published NC-Liverpool reference genome using a BLASTN search (version 2.7.1), to facilitate assignment of each transcript to a chromosome. High confidence hits (*E*-value $\leq 1E^{-50}$, Bit-score $\geq$ 200, and PID $\geq$ 90%) were subsequently cross-referenced with the contig location of each SNP, for allocation of SNPs to a chromosome. A BLASTX search was also performed querying the de novo NC-Liverpool transcriptome assembly against NC-Liverpool annotated proteins from ToxoDB (NcaninumLIV, version 30), to assign SNPs to a gene ID (PID $\geq$ 90%). The location of SNPs along the *N. caninum* genome was then

plotted to investigate the distribution of variants. Circos plots (Krzywinski et al. 2009) were generated to present the SNP data set in the context of the *N. caninum* genome, and to determine whether any particular regions might represent mutational (i.e., SNP) hotspots. A gene region was classified as a mutational hotspot if it contained >15 SNPs within a 50 kb window. The genes within these regions and their SNP information were extracted for gene ontology analysis.

Functional Analysis of Mutational Hotspots

InterProScan (version 68.0) was used to assign functional information to proteins putatively encoded by the genes within each SNP hotspot (Quevillon et al. 2005; Finn et al. 2017). For each of these hotspot genes, their respective protein sequences were analyzed using TMHMM (Sonnhammer et al. 1998), Phobius (Kall et al. 2004), and Philius (Reynolds et al. 2008), to identify potential transmembrane helices and signal peptides. The transcripts were also ranked by SNP density (contig length/number of SNPs) to investigate whether genomic regions with either a high or low SNP density were functionally significant. To do this, contigs were ranked on SNP density and contigs from within the first and third quartiles were extracted and their chromosome locations were identified. A *z*-test was performed to elucidate whether any chromosomes encoded a significantly larger number of SNP-dense contigs (P value <0.05). DNAPlotter (Carver et al. 2009) was then used to visualise the main features for chromosomes of interest, using available NC-Liverpool GenBank records (Ramaprasad et al. 2015).

Identifying Nonsynonymous and Synonymous Mutations

A de novo transcriptome was generated for NC-Nowra using Trinity, as described previously for NC-Liverpool. TransDecoder (Haas et al. 2013) was used to identify candidate protein-coding regions within the transcripts based on nucleotide composition and open reading frame (ORF) length. The protein sequences generated by TransDecoder for NC-Nowra were subjected to a BLASTP search (PID $\geq$80%) against the protein sequences generated for NC-Liverpool via the same procedure, to identify transcripts with identical ORFs, and those with mismatches or gaps between the two strains. These two lists were then cross referenced against the list of SNPs identified by VarScan, to identify nonsynonymous and synonymous mutations. The protein sequences from transcripts found to contain nonsynonymous SNPs were subsequently submitted to InterProScan for functional annotation, and the elucidation of gene ontologies, domains, repeats, and protein superfamilies.

Polymerase Chain Reaction (PCR) and Sanger Sequencing

PCR primers were designed to capture randomly selected variants identified by VarScan (supplementary file S1, tables S3 and S4, Supplementary Material online). All PCRs were prepared using the reagents provided in a MyTaq (Bioline) PCR kit. Each reaction contained 10 μM of each forward and reverse primer, 0.5 μl of MyTaq DNA Polymerase (5 U/μl), 2 μl of DNA template, and 5 μl of 5$\times$ MyTaq reaction buffer in a total volume of 50 μl. Each reaction was accompanied by a negative control, where DNA template was substituted with ddH_2O. The temperature cycling conditions employed were as follows: 1) 95 °C for 5 min, 2) 95 °C for 1 min, 3) 57–61 °C (primer dependent—see supplementary file S1, tables S3 and S4, Supplementary Material online) for 40 s, and 4) 72 °C for 40 s. Steps 2–4 were repeated 39 times, followed by a final extension step (5) of 72 °C for 5 min. PCR was performed on genomic DNA extracted from cultures of NC-Liverpool and NC-Nowra, as well as NC1, JPA1, NC-SweB1, WA-K9, NC-Beef, BPA1, BPA6, and an additional NC-Liverpool that had been cryogenically stored since 1998. This NC-Liverpool passage from 1998 was included as a control to investigate the genetic stability of this isolate over several years. The PCR products were then subject to electrophoresis on 2% agarose gels containing GelRed, visualized under UV light, and excised from gels using a sterile scalpel blade. Amplicons were purified from gel slices using a Qiagen QIAquick Gel Extraction Kit in accordance with the manufacturer's instructions. Sequencing was performed twice in both the forward and reverse direction on an ABI capillary sequencer, by the service provider Macrogen (South Korea). The ABI files were analyzed using SeqTrace (Stucky 2012). The forward and reverse sequences were assembled into contigs using an online version of CAP3 (Huang and Madan 1999). The resulting contigs were aligned for comparison using Clustal Omega (Sievers and Higgins 2014).

A summary of the workflow discussed above in its entirety is presented in figure 1, including the data sets exploited and created, the tools and software employed, and the analyses conducted.

Population Structure

To investigate whether an underlying population structure existed amongst the *N. caninum* strains studied based on identified sequence polymorphisms, a neighbor-joining tree was generated from a genetic distance matrix using the neighbor-joining tree estimation method of Saitou and Nei (1987). This was performed with the "nj" function within the "ape" R package, using the sequencing data generated through PCR analysis for each isolate as input.

DNA extracted from cultured *N. hughesi* tachyzoites was subjected to PCR amplification and sequencing of the same polymorphic loci, to investigate whether the confirmed variants identified in this study for *N. caninum* isolates were present. The sequences were used to generate an additional neighbor-joining tree incorporating the ten *N. caninum* strains, as well as *N. hughesi*.

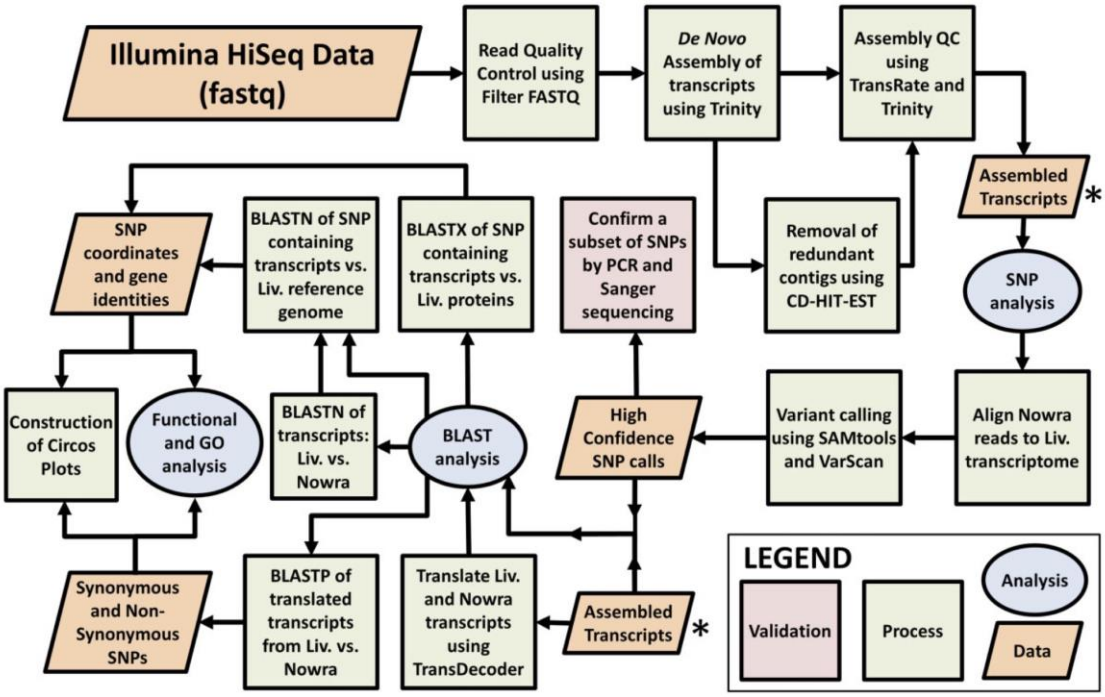

Fig. 1.—A summary of the complete variant detection workflow used in this study. Beginning with NGS (Illumina HiSeq) data, the reads generated from NC-Liverpool and NC-Nowra were used to assemble individual transcriptomes, and identify sequence variations between the two strains. A subset of high confidence variants was subsequently selected for laboratory validation for a total of ten *N. caninum* strains, each of which differs in pathogenic capability, geographical origin, and source. The SNP callset was then subjected to various computational analyses to determine their genomic location and functional significance, identify highly polymorphic regions, reveal whether their presence resulted in nonsynonymous (missense) or synonymous (silent) mutations, and determine an underlying population structure.

Comparison to *Toxoplasma gondii* Markers

Genetic markers commonly used in RFLP analyses of *T. gondii* were identified and cross referenced to the *N. caninum* SNP callset, in an attempt to elucidate whether the two closely related species were comparable with respect to genotyping markers (Lorenzi et al. 2016; Ruffolo et al. 2016). Additionally, based on the current understanding of genomic variation exhibited among the four major *T. gondii* lineages (Boyle et al. 2006; Khan et al. 2006; Khan et al. 2011b), the genes located on chromosome Ia in *N. caninum* were examined to see whether SNPs identified in this study mapped to this locus.

Results

Generation of a Reference Transcriptome

VarScan identified 1,520 high confidence SNPs following mapping of the NC-Liverpool transcriptome reads to the published NC-Liverpool reference genome. A total of 12 SNPs from this callset were subsequently confirmed through PCR analysis and Sanger sequencing as true differences between the two sources (see supplementary file S1, table S3, Supplementary Material online). The NC-Liverpool DNA sequenced from a cryopreserved 1998 NC-Liverpool culture (sourced from Liverpool University [Barber et al. 1993]) and the 2017 NC-Liverpool passage sequenced for this study were identical at these SNP locations. This confirmed the genetic stability of the strain over time, and indicated that the NC-Liverpool strain cultivated in-house was either genetically distinct from the published NC-Liverpool genome, or that the published NC-Liverpool genome contained some erroneous SNPs. This led us to use our in-house de novo transcriptome assembly as our reference for variant calling, given we could validate it by Sanger sequencing. Following removal of redundant contigs using CD-HIT-EST, 45,297 transcripts (27,570,740 assembled bases) remained in the Trinity assembly for use as a reference for mapping NC-Nowra reads and subsequent variant calling. A summary of the assembly metrics is contained within supplementary file S1 and S2, Supplementary Material online.

Variant Calling

The NC-Nowra and the NC-Liverpool tachyzoite transcriptomes differed by 3,130 SNPs and 6,123 indels (table 1). Supplementary file S3, Supplementary Material online, contains a list of NC-Liverpool transcriptome contigs containing SNPs identified by VarScan, along with the variant positions, and reference and alternate bases. The number of SNPs observed in a contig ranged from 0 to 28, the average being 1.55 SNPs per contig. A total of 1,838 transitions (A/G and C/T) and 1,292 transversions (A/C, A/T, C/G, and G/T) were observed between the NC-Nowra and NC-Liverpool strains, representing a transition/transversion ratio (T_i/T_v) of 1.42. The 3,130 high confidence SNPs were distributed across 2,021 unique transcripts encoded by 1,879 genes. Additionally, the current *N. caninum* reference genome consists of multiple large contigs that are not assigned to one of the 14 chromosomes. There were 162 SNPs distributed across 22 such contigs, with the majority of this callset (~78%) assigned to eight of these contigs alone. These SNPs were subsequently allocated to 34 unique protein-coding genes. It is worth noting that when blasting the NC-Liverpool transcriptome against the published *N. caninum* annotated proteins to assign SNPs to annotated genes, several contigs containing SNPs returned high confidence BLAST hits (i.e., PID $\geq$ 90%) to multiple genes along the genome, however not all SNPs were assigned to protein-coding genes based on the BLAST results.

A set of 27 variants identified in silico between NC-Liverpool and NC-Nowra were subject to PCR and sequencing analysis (table 2). No false-positive variants were identified from among the 27 variants examined, though the workflow failed to detect three true variants (i.e., three false negative SNPs) within these genetic markers, as revealed by Sanger sequencing. It was found that VarScan originally identified these variants, but they were discarded during subsequent filtering steps.

Table 1

Summary of VarScan Variant Calling Using NC-Nowra RNA-seq Reads Aligned to the *De Novo* NC-Liverpool Transcriptome Reference

	Number of SNPs Called		Number of Indels Called	
	Prefiltering	Postfiltering	Prefiltering	Postfiltering
Pre-CD-HIT-EST	15,807	3,562	8,163	5,067
Post-CD-HIT-EST	15,361	3,130	8,966	6,123

Note.—Most SNPs initially called by VarScan were discarded following filtering based on strand bias, sequence, and variant coverage thresholds, mismatch quality sum, and read position bias.

Table 2

A Summary of the Total Number of Variants Selected for and Confirmed Through Sequencing, Based on Targeting Various Loci

Metric	Value
High confidence SNPs called by VarScan between NC-Liverpool transcriptome and NC-Nowra reads	3,130
Variants captured and confirmed through Sanger sequencing	37
Variants captured and confirmed in MLST targets	27
Variants captured in MLST, identified as false positives through Sanger sequencing	0
Variants discovered via Sanger sequencing, identified as false negatives by VarScan	3
Total variants captured in MLST, sequenced for a total of ten *N. caninum* strains	30

Distribution and Functional Annotation of SNPs

Multiple SNP hotspots were identified, distributed unevenly throughout the *N. caninum* genome (fig. 2). A large number of SNPs clustered on chromosome XI (FR823392), in addition to various hotspots identified in chromosomes V (FR923386), VI (FR823387), and XII (FR823393). There were 19 hotspots containing 73 *N. caninum* genes, many of which were implicated in translation (*NCLIV_057380* and *NCLIV_057360*), transcription (*NCLIV_057870* and *NCLIV_065940*), ribosomal subunit formation (*NCLIV_056680*, *NCLIV_056820*, *NCLIV_056830*, and *NCLIV_057070*), GTP binding and GTPase activity (*NCLIV_057820* and *NCLIV_057390*), protein transport (*NCLIV_057490*), and kinase activity or protein phosphorylation (*NCLIV_ 056620* and *NCLIV_057940*). The genomic location of these 73 genes contained within SNP hotspots and their annotations are tabulated in supplementary file S1, table S5, Supplementary Material online. Thirty-five genes from this callset contained five or more SNPs.

Prior to functional analysis, 23 of the 35 SNP hotspot genes were identified as hypothetical proteins or unspecified products based on their corresponding gene IDs. All but nine were assigned GO terms, protein families, domains and/or repeats by InterProScan. Protein superfamilies that appeared more than once among these 35 SNP hotspot genes included WD40 repeat containing domain superfamily (IPR036322), ARM-like helical (IPR011989), ARM-type fold (IPR016024), and P-loop-containing nucleoside triphosphate hydrolase (IPR027417). Other superfamilies of functional interest were zinc finger RING/FYVE/PHD type (IPR013083), Sec1-like superfamily (IPR036045), EF-hand domain pair (IPR011992), ABC transporter superfamily (IPR036640), and the translation initiation factor eIF-4e-like (IPR023398) superfamilies. Domains and repeats featured were AAA+ ATPase domain (IPR003593), tetratricopeptide repeat (IPR019734), subtilisin SUB1-like catalytic domain (IPR034204), and WD40-repeat-containing protein (IPR017986).

Recurring Gene Ontologies (GO) for molecular function included protein binding (GO: 0005515), binding (GO: 0005488), and hydrolase activity (GO: 0016787). Regarding biological process GOs, those assigned included lipid metabolic process (GO: 0006629), translation initiation

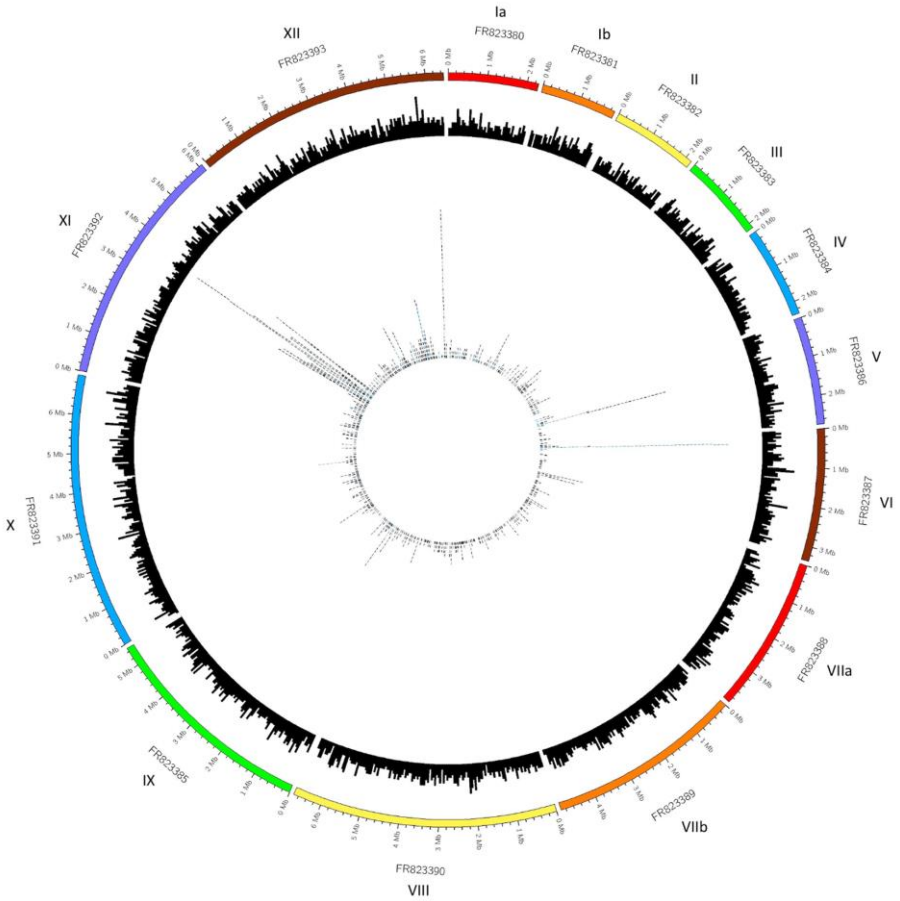

Fig. 2.—Circos plot representing the SNP data in the context of the *N. caninum* genome. The outer track is an ideogram representing the 14 *N. caninum* chromosomes and their sizes, followed by a histogram of the 7,121 annotated genes along each chromosome in the middle track. This histogram is based on each gene's location within a chromosome, plotted in 50 kb windows, relative to the ideogram. The inner most track contains the distribution of identified SNPs as located within these annotated genes. Each tile in this inner track represents a SNP that has fallen within that gene region, relative to the ideogram. The tiles are also colored based on their size, where those that are blue represent genes that are larger than 10,000 bases.

(GO: 0006413), metabolic process (GO: 0008152), proteolysis (GO: 0006508), and transmembrane transport (GO: 0055085). Supplementary table S6 within supplementary file S1, Supplementary Material online contains a complete list of gene annotation information and ontologies for the putative proteins encoded within these SNP hotspots.

The three bioinformatic tools employed to identify transmembrane (TM) proteins and signal peptides within the SNP hotspot list, did not present consistent results for all protein sequences explored. However, mutually reported between both Phobius and Philius, were four transmembrane proteins, and six globular proteins with signal peptides, all of which were present on either chromosome VI or XI, except for one signal peptide containing protein which was located on chromosome V. Two hotspot genes also encoded transmembrane proteins with signal peptides, both of which were located on chromosome XI (*NCLIV_056900* and *NCLIV_057550*). Interestingly, two TM proteins and two signal peptide containing proteins could not be assigned any additional annotations or gene ontologies.

Estimation of Synonymous and Nonsynonymous SNP Count

When the translated transcriptomes of NC-Nowra and NC-Liverpool were compared, 652 SNPs were found to be located in open reading frames that possessed different translations between the strains, and these SNPs were distributed across 287 unique genes. There were also 470 SNPs assigned to a protein-coding gene where the translation of the respective transcript was identical between the strains (i.e., synonymous mutations). However, where the number of mismatches reported by BLASTP exceeded the number of SNPs within a contig, it was assumed that VarScan had filtered out real

sequence variants between the two strains (i.e., false negatives SNPs). Alternatively, in the event that there were more SNPs identified by VarScan within a contig than BLASTP mismatches, the additional SNPs were presumed to result in synonymous mutations. Therefore, it was estimated that the final VarScan SNP callset contained at least 468 nonsynonymous SNPs, and 654 synonymous SNPs. Figure 3 displays the distribution of the nonsynonymous and synonymous SNPs identified across the *N. caninum* genome. Many nonsynonymous mutations coincided with the locations of the SNP hotspots identified, including those on chromosomes VI (FR823387), XI (FR823392), and XII (FR823393), whereas almost all the SNPs located on chromosome V (FR823386) were found to be synonymous mutations. Additionally, the aforementioned callsets included 60 nonsynonymous SNPs, and 63 synonymous SNPs part of large contigs within the *N. caninum* genome, which are not pictured in the Circos plots generated.

Functional Analysis of Transcripts Containing nonsynonymous SNPs

The GOs that were overrepresented in transcripts containing nonsynonymous SNPs from amongst the molecular function GO category included protein kinase activity (GO: 0004672), ATP binding (GO: 0005524), and protein binding (GO: 0005515). Recurring GOs from the biological process category included protein phosphorylation (GO: 0006468), proteolysis (GO: 0005576), and oxidation–reduction process (GO: 0055114). The protein superfamilies repeatedly featured were protein kinase-like domain superfamily (IPR011009), p-loop containing domain-like superfamily (IPR027417), WD40-repeat containing domain superfamily (IPR036322), and tetratricopeptide-like helical domain superfamily (IPR011990). Recurring protein domains of functional importance included protein kinase (IPR000719), AAA+ ATPase (IPR003593), EF-hand calcium binding (IPR018247), and PAN/Apple domain (IPR003609), as well as featured protein repeats such as WD40 repeat (IPR001680), and Ankyrin repeat (IPR002110). Also of interest as reported by InterProScan, were protein signatures such as serine/threonine protein kinase active-site signature, protein kinase ATP binding site signature, protozoan surface antigen signature (SAG1), and ABC transporters family signature.

Twenty-seven of the 35 SNP hotspot genes were found to contain nonsynonymous SNPs, including proteins coding for kinesin, SUB2, an ABC transporter, a Sec1 protein, and fatty acyl-CoA desaturase.

Distribution of Transcripts of High and Low SNP Densities

Chromosome XI (FR823392) possessed the largest number of contigs with a high SNP density across the genome. The *z*-test confirmed that the two chromosomes encoding a significantly larger number of SNP-dense transcripts (P value <0.05), compared with the number of contigs with a low SNP density, were chromosomes VI (FR823387) and XI (FR823392). Figure 4*A* and *B* depict the main genomic features of chromosomes VI and XI, plotted using the available GenBank records for chromosome VI (LN714480.1) and XI (LN714480.1). Both chromosomes are transcriptionally active in *N. caninum* tachyzoites, and only a very small number of noncoding regions exist between genes. The SNP hotspots within these chromosomes seem localized to selected genomic windows. The chromosomes also encode ncRNA (noncoding RNA) molecules, including tRNAs, dispersed unevenly along the length of each chromosome. Additionally, there are clear areas where the GC content along the chromosome either peaks above average, or decreases. Some SNP hotspots on either chromosome also appear to coincide with peaks in GC content, such as that on chromosome VI between approximately 450,000 and 500,000 bases.

Genetic Population Structure

The Sanger data generated for eight additional *N. caninum* strains, across nine selected loci (NC1, JPA1, NC-SweB1, WA-K9, NC-Beef, BPA1, BPA6, and an additional NC-Liverpool strain cryogenically frozen since 1998) containing 30 of the confirmed variants, did not reveal any specific patterns of segregation (i.e., geographical or otherwise). Nevertheless, JPA1, BPA1, and NC1 were more similar to NC-Liverpool, whereas NC-Nowra, NC-SweB1, NC-Beef, BPA6, and WA-K9 were more similar to each other than to the formerly mentioned isolates. The NC-Liverpool strains from different passage numbers were identical. The neighbor-joining tree presented in figure 5*A* revealed the grouping of the ten strains into two distinct clades based on the SNP data: the virulent strains including NC-Liverpool, and the more attenuated group including NC-Nowra and NC-SweB1.

Seven of the nine sequenced polymorphic loci (supplementary file S1, table S4, entries 1–7, Supplementary Material online) which contained 28 confirmed variants between NC-Liverpool and NC-Nowra identified in this study, were able to be amplified for *N. hughesi*, through the DNA extraction and PCR amplification methodology described. The sequencing results for *N. hughesi* revealed the presence of all 28 variants identified in silico for *N. caninum*, as well as the existence of an additional 28 SNPs that were unique to *N. hughesi*. Presented in figure 5*B* is an additional neighbor-joining tree generated from seven of the nine aligned polymorphic loci, for *N. hughesi* and all ten *N. caninum* strains. These results provide further additional support for the existence of two clades of *N. caninum*.

Comparison to *Toxoplasma gondii* Markers

Of the 12 *T. gondii* genotyping markers examined, only two SNPs were present within *N. caninum* orthologous genes.

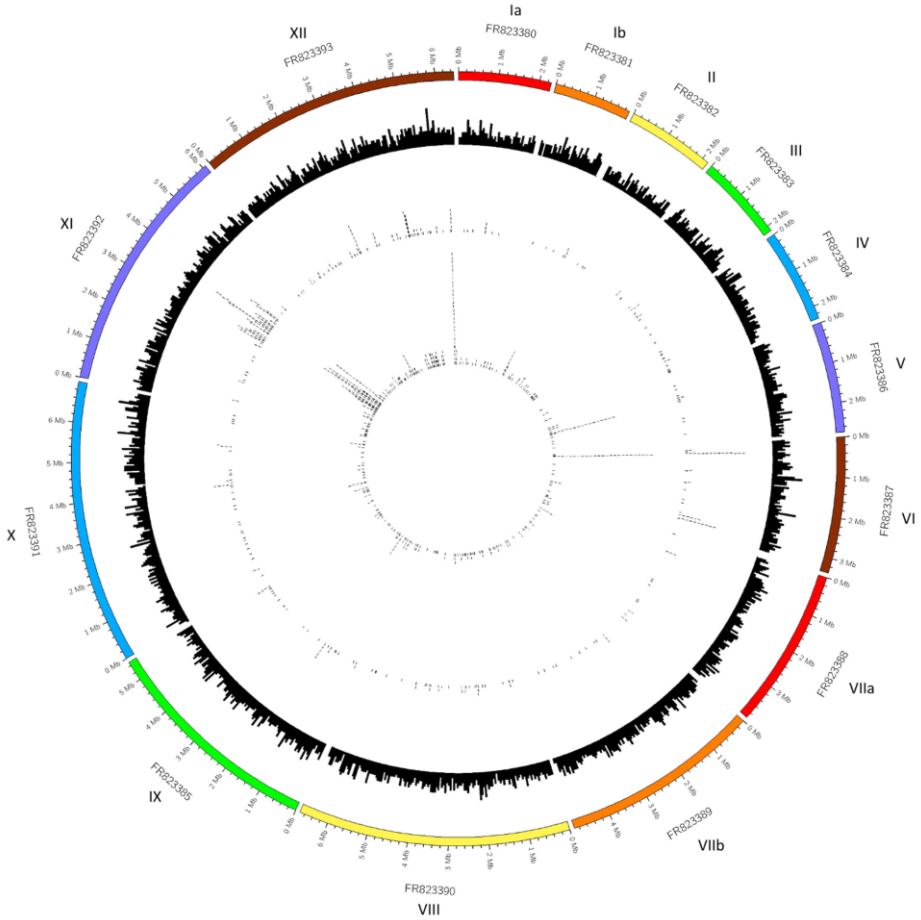

Fig. 3.—Circos plot presenting the location of nonsynonymous SNPs identified in the *N. caninum* genome. The outer track is an ideogram, representing the 14 *N. caninum* chromosomes and their sizes, followed by a histogram of the 7,121 annotated genes along each chromosome on the second track from the outside. This histogram is based on each gene's location within a chromosome, plotted in 50 kb windows, relative to the ideogram. The next track (third from the outside) represents the locations of nonsynonymous SNPs called by VarScan within these annotated genes. Similarly, the innermost track depicts the locations of synonymous SNPs identified by VarScan.

They were in dense granule protein (*GRA7*) and class I chitinase (*CLP1*). These results appear to be consistent with that of Al-Qassab et al. (2010b), where no sequence differences were detected in various proteins of canine and bovine *N. caninum* strains. This included *SAG1*, *SRS2*, and *GRA6*, all of which are among the 12 *T. gondii* genotyping markers explored in this study. Due to the location of one SNP at the very beginning of the contig, PCR and sequencing analysis was only performed and confirmed for one of these SNPs, as per the last entry in supplementary table S4 of supplementary file S1, Supplementary Material online.

After assigning the location of each SNP to a chromosome (fig. 2) to examine the distribution of variation along the *N. caninum* genome, it was observed that chromosome Ia (FR823380) had the second lowest SNP density with <100 SNPs being present, only second to chromosome Ib (FR823381).

Discussion

Neospora caninum is an apicomplexan parasite, responsible for reproductive failure in cattle and neurological disease in dogs. Intraspecies diversity is known in the form of extreme differences in virulence between strains found across the globe. The genetic basis of this diversity is unknown though an improved understanding of this could help to identify novel virulence loci.

We used a bioinformatics workflow to identify genome-wide genetic differences between two phenotypically distinct strains of *N. caninum*. These strains vary drastically in their pathogenic propensity, and represent extremes of *N. caninum* virulence (Atkinson et al. 1999; Miller et al. 2002). A variant analysis workflow was employed to identify SNPs present in the genomes of NC-Liverpool and NC-Nowra, and the SNPs were subsequently subjected to laboratory validation through PCR and Sanger sequencing. A multilocus sequencing

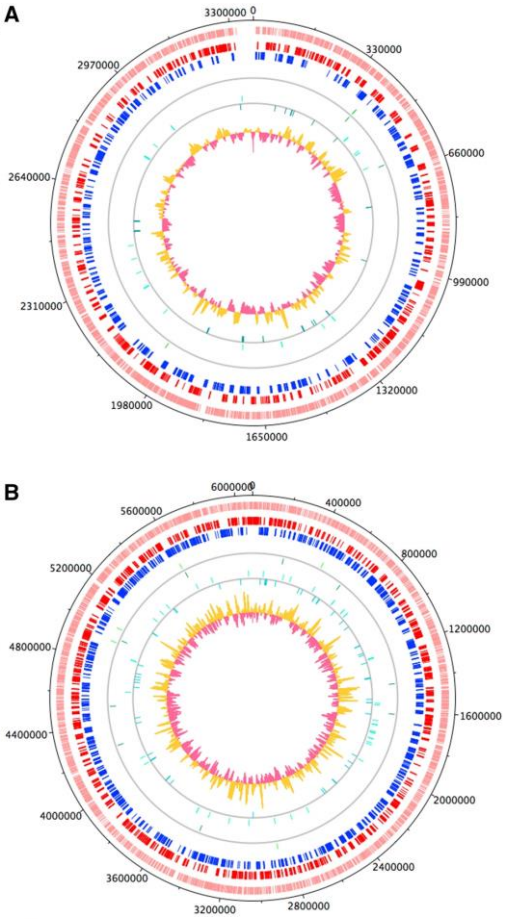

Fig. 4.—Plots displaying the features of chromosome VI (*A*) and XI (*B*). The outer most track (pink) in both plots represents the location of all genes on either chromosome. The adjacent two tracks contain the CDS on the forward (red) and reverse (blue) strands. The green strokes on the next two tracks show the location of tRNAs on the forward and reverse strands, respectively, and similarly the location of ncRNAs are represented by the aqua strokes on either strand. The GC content along the chromosome is displayed in the second most inner track, where the yellow depicts areas above average content, and the pink being below average.

approach was developed using this information, comprising of nine randomly selected loci, with a combined length of 3.4 kb and containing 30 validated variants. This method was applied to ten *N. caninum* strains, including two NC-Liverpool samples from different passages and NC-Nowra, to reveal a population structure consisting of two major clades. We also identified SNP hotspots within the genome of *N. caninum*, characterized by elevated levels of SNP density.

The choice of variant caller for genome-wide SNP detection requires careful consideration and optimization depending on the organism under investigation and the data available (Reumers et al. 2012; O'Rawe et al. 2013; Bao et al. 2014; Pabinger et al. 2014; Ribeiro et al. 2015). In addition, some variant callers such as the routinely used Genome Analysis Toolkit (GATK) (McKenna et al. 2010) require a database of known SNPs to preprocess reads for variant calling, and consequently fail to consider nonmodel organisms such as *N. caninum*. As the majority of variant callers are like the GATK and depend on reference-based mapping, their use in nonmodel species is often restricted due to the absence of high-quality reference genomes (Dou et al. 2012).

The VarScan package employed in this study exploits empirical and statistical thresholds based on user-defined criteria to call variants, representing a simple pipeline that is compatible with several short-read aligners (Koboldt et al. 2013). This versatility means it can be applied to nonmodel organisms such as *N. caninum*. Using its default recommended parameters, the VarScan 2 pipeline (Koboldt et al. 2013) identified thousands of high confidence SNPs and indels between the de novo NC-Liverpool transcriptome generated in-house, and the NC-Nowra RNA-seq data. Of the hundreds of variants that were randomly selected for manual visualization in IGV, most exhibited robust quality scores and high sequence coverage at the variant position. Sanger sequencing identified a small number of false negative variants filtered out subsequent to variant calling. This highlights the need for careful optimization of filtering parameters and the necessity of validating SNPs identified *in silico* by Sanger sequencing before deriving any biological conclusions.

Ribeiro et al. (2015) explored the relationship between the choice of tools and parameters, and their impact on false positive variants. Out of the seven factors explored, the quality of the reference sequence used had the most pronounced effect on the false positive variant calling rate. This finding raises concerns for the use of similar variant calling pipelines on nonmodel organisms in the early stages of genomic examination where the reference genomes may be poor or misassembled, the product of limited or incomplete sequencing, or the result of inadequate quality control and validation. This can subsequently result in errors in the reference sequence being identified as read mismatches, producing false positive variants.

Using the NC-Liverpool genome from ToxoDB as a reference, all the variants called were false positives, typically located at the ends of reads or in homopolymer runs, which are known error sources associated with DNA sequencing (Reumers et al. 2012; Durtschi et al. 2013). This discovery prompted the assembly of a de novo transcriptome using RNA-seq data derived from the NC-Liverpool parasites cultured in-house. The variant calling workflow employed here identified numerous SNPs when comparing our cultured NC-Liverpool strain to the ToxoDB NC-Liverpool reference,

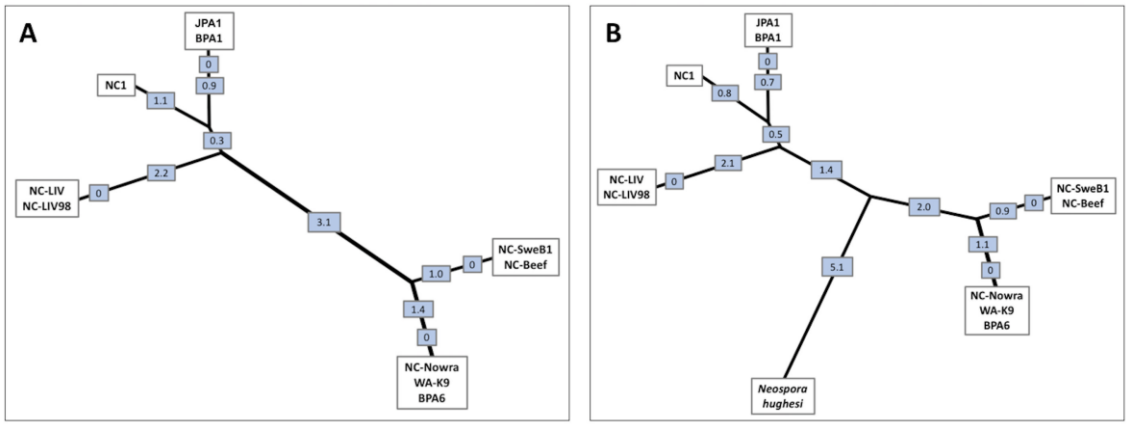

Fig. 5.—Unrooted neighbor-joining trees showing the population structure within the *Neospora* genus. (*A*) This unrooted neighbor-joining tree was generated from pairwise genetic distances calculated from nine genetic markers capturing 30 variants, including ten *N. caninum* strains (including one repeat of NC-Liverpool; NC-LIV98). The tree suggests a population genetic structure comprising two major lineages of *N. caninum*. (*B*) This neighbor-joining tree was generated from seven of the nine genetic markers capturing 28 confirmed variants for *N. caninum* strains, as well as an additional 28 SNPs that were unique only within the *N. hughesi* sequences. The values displayed in both trees indicate the genetic distance between nodes.

suggesting the reference in ToxoDB was erroneous. Although alternatively it is possible that such identified variants could represent allelic variation between these NC-Liverpool cultures, this does not seem to be the case, at least for the set of 12 variants that were confirmed through sequencing. Based on our extensive work and the original use of the ToxoDB NC-Liverpool genome as a reference for variant calling, the confirmed SNPs are most likely attributable to errors in this published genome. The absence of any differences between the 2017 and 1998 NC-Liverpool strains based on our MLST approach also supports this.

Analysis of the distribution of identified SNPs elucidated the existence of SNP hotspots across the *N. caninum* genome (fig. 2), especially their clustering on chromosome VI (FR923387), XI (FR923392), and XII (FR823393). However, the current (or absence of) annotation of the *N. caninum* genome presented a problem for assigning functional significance to the SNPs identified in this study, and more broadly remains a problem for the study of virulence and pathogenicity within the species. The fact that 4,011 of 6,936 genes in the published *N. caninum* genome are annotated as hypothetical proteins, presents a major and concerning hindrance to the study of potential virulence factors. Furthermore, recent studies focusing on improving and expanding the available gene structure and annotations for *N. caninum* are yet to appear in ToxoDB reference resources (Goodswen et al. 2015; Krishna et al. 2015; Ramaprasad et al. 2015). Although 3,130 high confidence SNPs were called and 19 genomic SNP hotspots identified, many were located within the coding regions of hypothetical proteins or uncharacterized genomic regions, which greatly hindered the ability to assign biological context to these polymorphic regions.

In an effort to annotate the corresponding protein sequences for each SNP hotspot identified in this study, many of which were hypothetical proteins, various tools such as InterProScan were used. Within these hotspots were two genes coding for WD40 domain containing protein: *NCLIV_057900* and *NCLIV_013170*. WD40 repeat containing proteins belong to one of the largest, most abundant protein families found in all eukaryotes (Neer et al. 1994). These proteins are associated with a variety of functions including signal transduction and transcription regulation, cell cycle control, autophagy, apoptosis, transmembrane signaling, and cytoskeleton assembly. The fundamental shared function of all WD40-repeat proteins is facilitating multi-protein complexes, where the repeats serve as a rigid scaffold for protein interactions. The significance of this is that for intracellular protozoan parasites, the efficiency of infection is contingent on the parasite's capacity for host cell recognition, adhesion, and invasion, which are generally mediated by protein–protein interactions (von Bohl et al. 2015).

InterProScan characterized one hypothetical, SNP hotspot protein (*NCLIV_057320*) as belonging to the tetratricopeptide-like helical domain superfamily. As with members of the WD40 family, TPR containing proteins are involved in protein–protein interactions and various metabolic and regulatory processes, and thus play an important role in virulence (Goebl and Yanagida 1991). Also amongst the most abundant proteins in eukaryotes, and characterizing one identified hotspot gene, zinc finger domain containing proteins exhibit versatile binding modes, suggesting that such motifs are stable scaffolds with specialized functions. Zinc finger proteins are involved in transcription and translation regulation, DNA

and RNA recognition, protein folding and assembly, apoptosis, and cell adhesion (Laity et al. 2001).

The ATP binding cassette (ABC) superfamily of proteins are expressed as efflux transporters in eukaryotes, that translocate a plethora of substrates including proteins, ions, toxins and amino acids across membranes (El-Awady et al. 2017). All ABC transporters consist of two domains: the nucleotide binding domain (NBD) and the (transmembrane domain [TMD], where the coupling of these domains facilitates import and export). The protein hotspot identified as an ABC transporter, *NCLIV_065950*, had gene ontologies related to transmembrane transport (GO: 0055085), ATP binding (GO: 0005524), and ATPase activity coupled to transmembrane movement of substances (GO: 0042626). However, while Phobius recognized the transmembrane topology of this protein coding sequence, Philius and TMHMM did not.

Since the data exploited in this study was generated from RNA-seq data, it was unexpected that SNPs were identified that were not located in annotated genes. This suggests that either the current gene annotation is incorrect or incomplete, or that new/novel abundantly expressed transcripts were present in the culture from which the RNA-seq data was generated. However, this study did not investigate the presence of sequence variants located within apicoplast or mitochondrial DNA, to which some of the identified SNPs may have been located within.

In addition to the mutational hotspots revealed throughout the *N. caninum* genome in this study, the nonsynonymous mutations identified can also contribute to the current understanding of pathogenic variability within the species. As a nonsynonymous SNP alters a protein's sequence, their presence can cause changes in biochemical activity, protein–protein interactions, and molecular function, which can consequently establish the link between genotype and biologically significant phenotypes (Ng and Henikoff 2006; Zhao et al. 2014; Tang and Thomas 2016). This stresses the importance of not only identifying and comparing sequence variants present between populations, but also understanding whether such mutations have the potential to disrupt the resulting protein's function. The identification of nonsynonymous SNPs within protein coding genes in this study may provide new insight into and sources for studying the underlying causes of phenotypic differences between isolates of *N. caninum*, presenting new potential determinants of virulence and pathogenic capability.

Analyzing and recognizing the existence of population structure within a species is conducive to understanding and determining the spread of virulence factors within and between geographic locations (Khan et al. 2011b). As presented in figure 5*A*, the ten strains, including two NC-Liverpool strains from different passages, comprise two distinct genetic clusters that may reflect differences in pathogenicity. The highly virulent NC-Liverpool strain was the most distinct type, and was placed at a genetic distance furthest from the clades containing the less virulent NC-Nowra and NC-SweB1 strains, but at a small distance from the virulent NC1 strain. Although significant differences in virulence between select *N. caninum* strains have been published in either mice or cattle models (Dubey et al. 1992; Atkinson et al. 1999; Innes et al. 2001; Miller et al. 2002; Maley et al. 2003), limited studies currently exist that comprehensively document the pathogenic variability of many other isolated strains, including NC-Beef, BPA6, and WA-K9. This makes it difficult to corroborate the population structure elucidated in this study and make an assumption regarding virulence, based on the presence or absence of sequence variations investigated. However, the neighbor-joining tree presented in figure 5*B* with the inclusion of *N. hughesi*, supports the existence of a two-clade population structure for *N. caninum*, dividing the ten strains into genetic clusters potentially resembling their virulence properties. We refrain from suggesting that *N. caninum* as we know it, may represent two independent species. The relationship represented in figure 5*B* including *N. hughesi* suggests that this idea is worth investigating further. It is also worth mentioning that the two clades elucidated in this study reflect the results of the Tand-12 minisatellite marker described by Al-Qassab et al. (2010a) for these isolates. The NC-Liverpool cluster contained three copies of this repeat, whereas the NC-Nowra cluster is characterized by four copies of this repeat.

Fatality was observed in only one of eight susceptible γ-INF-KO mice infected with NC-Beef oocysts, Lindsay et al. (1999) suggested that this strain may be characterized by a lack of pathogenicity. Additionally, WA-K9 was the first canine strain from Australia, cultivated from skin lesions found on a dog in Western Australia (McInnes et al. 2006). What was noteworthy about the clinical presentation of this dog was that infection initially manifested as cutaneous neosporosis, where the parasite is primarily responsible for neurological illness in canines. However, the dog was essentially normal at a 2.5 year follow up examination after continuous treatment with a high dosage of clindamycin, and subsequent to initial treatment and recrudescent infection. The successful treatment and opportunistic infection characterizing this particular case, may suggest reduced virulence of this strain, and hence further affirm the population structure determined in this study.

The NC-Liverpool DNA sequenced from the 1998 culture (sourced from Liverpool University [Barber et al. 1993]) and the 2017 passage sequenced for this study were identical at the genomic locations studied in MLST, confirming the genetic stability of the strain over time, and indicating that the NC-Liverpool cultivated in-house was either genetically distinct from the published NC-Liverpool genome, or that this genome contains erroneous SNPs. It is also important to note that the 1998 isolate is known to be virulent in mice (Atkinson et al. 1999). Additionally, the absence of virulence in the NC-Nowra isolate was confirmed as recently as 2013 in cattle vaccine trials (Weber et al. 2013). However, it should be noted

that this study did not compare other isolates previously categorized as virulent in cattle, such as NC-Liverpool.

Although studies have established varying degrees of intraspecies genetic diversity within *N. caninum* in repeat regions, it is expected that SNPs will replace repetitive sequences as DNA markers, due to their distribution throughout the entire genome and their low mutation rates (Picoult-Newberg et al. 1999). As variant identification using RNA-seq data from *N. caninum* is unprecedented, it is valuable to compare these results to genetic variation identified in well-studied model organisms. *Toxoplasma gondii* is a model Apicomplexan with robust data available, and has been thoroughly studied to elucidate existing genotypes, population structure, and potential virulence markers. The plethora of studies exploring the population structure of *T. gondii* has shown that the global between-lineage variation ranges from approximately 0.01 to 5% (Boyle et al. 2006). It is well documented that a distinct split exists between *T. gondii* lineages found in North America and Europe, compared with those in South America (Khan et al. 2011b). Furthermore, the cause of most infections in the Northern Hemisphere can be traced to four clonal lineages, each with differing levels of pathogenicity (Khan et al. 2011a). It was observed that very few sequence polymorphisms exist on chromosome 1a between these dominant lineages (Khan et al. 2006, 2011b). Due to this common monomorphic chromosome, the current model of evolution suggests that approximately 10,000 years ago a genetic sweep caused the expansion of these lineages, from only a limited number of genetic crosses between highly related precursor strains (Boyle et al. 2006). After assigning the location of each SNP to a chromosome (fig. 2), it was observed that chromosome Ia in *N. caninum* had the second lowest SNP density with <100 SNPs being present. Whether *N. caninum* experienced a similar genetic sweep to *T. gondii* at the time is not entirely clear, however if this was the case, based on the existence of SNPs across this locus, such a sweep may not have been as severe.

In summary, this study shows that variant analysis can contribute to our understanding of the existence and underlying mechanisms of genetic diversity within the *N. caninum* species, as well as the mechanisms of virulence and pathogenesis. Based on this, SNP identification has the potential to replace mini- and microsatellite markers for exploring such intraspecies diversity. The MLST approach developed in this study reveals a population structure reflecting two major clades that do not support any obvious geographical segregation. This knowledge will facilitate the future identification of novel virulence markers and guide the selection of candidate components for a subunit vaccine against bovine neosporosis.

In addition, we present a bioinformatic workflow that identified thousands of genetic variants in loci that are transcriptionally active during the tachyzoite stage of the *N. caninum* life cycle. This data informed the development of an MLST approach based on nine transcriptionally active tachyzoite-associated loci that provides new insights on the population genetic structure of *N. caninum*. We also identify a set of *N. caninum* proteins as potentially novel virulence determinants for downstream investigation, based on both the presence of SNP-dense regions (hotspots), and nonsynonymous mutations within protein-coding genes. This work provides new insights into the molecular basis behind the marked virulence properties reported between strains of *N. caninum*, which is knowledge that will be pertinent to the future development of a subunit vaccine against bovine neosporosis.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Martin Krzywinski for his assistance with the Circos software. This study was completed by L.C. in partial fulfilment of the Ph.D. degree at UTS.

Literature Cited

Al-Qassab S, Reichel MP, Ellis J. 2010. A second generation multiplex PCR for typing strains of *Neospora caninum* using six DNA targets. Mol Cell Probes 24(1):20–26.

Al-Qassab SE, Reichel MP, Ellis J. 2010. On the biological and genetic diversity in *Neospora caninum*. Diversity 2(3):411.

Atkinson R, Harper PA, Ryce C, Morrison DA, Ellis JT. 1999. Comparison of the biological characteristics of two isolates of *Neospora caninum*. Parasitology 118(4):363–370.

Bao R, et al. 2014. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform. 13(2 Suppl):67–82.

Barber J, Trees A, Owen M, Tennant B. 1993. Isolation of *Neospora caninum* from a British dog. Vet Rec. 133(21):531–532.

Barr BC, et al. 1994. Experimental reproduction of bovine fetal Neospora infection and death with a bovine Neospora isolate. J Vet Diagn Invest. 6(2):207–215.

Basso W, et al. 2009. Molecular comparison of *Neospora caninum* oocyst isolates from naturally infected dogs with cell culture-derived tachyzoites of the same isolates using nested polymerase chain reaction to amplify microsatellite markers. Vet Parasitol. 160(1–2):43–50.

Blankenberg D, et al. 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics 26(14):1783–1785.

Blankenberg D, et al. 2010b. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. Chapter 19: Unit 19. 10. 1–21.

Boyle JP, et al. 2006. Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like *Toxoplasma gondii*. Proc Natl Acad Sci U S A. 103(27):10514–10519.

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. 2009. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25(1):119–120.

Caspe SG, et al. 2012. The *Neospora caninum*-Spain 7 isolate induces placental damage, fetal death and abortion in cattle when inoculated in early gestation. Vet Parasitol. 189(2–4):171–181.

Cerede O, et al. 2005. Synergistic role of micronemal proteins in *Toxoplasma gondii* virulence. J Exp Med. 201(3):453–463.

Dou J, et al. 2012. Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. Biol Direct. 7(1):17.

Dubey JP. 1999. Recent advances in Neospora and neosporosis. Vet Parasitol. 84(3-4):349–367.
Dubey JP, Dubey J. 2003. Review of *Neospora caninum* and neosporosis in animals. Korean J Parasitol. 41(1):1–16.
Dubey JP, Buxton D, Wouda W. 2006. Pathogenesis of bovine neosporosis. J Comp Pathol. 134(4):267–289.
Dubey JP, Carpenter JL, Speer CA, Topper MJ, Uggla A. 1988. Newly recognized fatal protozoan disease of dogs. J Am Vet Med Assoc. 192(9):1269–1285.
Dubey JP, Hattel AL, Lindsay DS, Topper MJ. 1988b. Neonatal *Neospora caninum* infection in dogs: isolation of the causative agent and experimental transmission. J Am Vet Med Assoc. 193:1259–1263.
Dubey JP, Lindsay DS. 1996. A review of *Neospora caninum* and neosporosis. Vet Parasitol. 67(1–2):1–59.
Dubey JP, Lindsay DS, Anderson ML, Davis SW, Shen SK. 1992. Induced transplacental transmission of *Neospora caninum* in cattle. J Am Vet Med Assoc. 201:709–713.
Durtschi J, Margraf RL, Coonrod EM, Mallempati KC, Voelkerding KV. 2013. VarBin, a novel method for classifying true and false positive variants in NGS data. BMC Bioinformatics 14(13 Suppl):S2.
El-Awady R, et al. 2017. The role of eukaryotic and prokaryotic ABC transporter family in failure of chemotherapy. Front Pharmacol. 7:535
English ED, Adomako-Ankomah Y, Boyle JP. 2015. Secreted effectors in *Toxoplasma gondii* and related species: determinants of host range and pathogenesis?. Parasite Immunol. 37(3):127–140.
Finn RD, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45(D1):D190–D199.
Goebl M, Yanagida M. 1991. The TPR snap helix: a novel protein repeat motif from mitosis to transcription. Trends Biochem Sci. 16(5):173–177.
Goodswen SJ, Barratt JL, Kennedy PJ, Ellis JT. 2015. Improving the gene structure annotation of the apicomplexan parasite *Neospora caninum* fulfils a vital requirement towards an in silico-derived vaccine. Int J Parasitol. 45(5):305–318.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.
Haas BJ, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494–1512.
Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9(9):868–877.
Huynh MH, Carruthers VB. 2006. Toxoplasma *MIC2* is a major determinant of invasion and virulence. PLoS Pathog. 2(8):e84.
Innes EA, et al. 2001. Protection against vertical transmission in bovine neosporosis. Int J Parasitol 31(13):1523–1534.
Recommended methods and typical workflows for multi-step analyses [Internet]. 2015. Available from: https://www.broadinstitute.org/gatk/guide/topic? name=methods; Last accessed January 2018
Kall L, Krogh A, Sonnhammer EL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338(5):1027–1036.
Keller N, et al. 2002. Identification of a *Neospora caninum* microneme protein (*NcMIC1*) which interacts with sulfated host cell surface glycosaminoglycans. Infect Immun. 70(6):3187–3198.
Khan A, et al. 2006. Common inheritance of chromosome Ia associated with clonal expansion of *Toxoplasma gondii*. Genome Res. 16(9):1119–1125.
Khan A, et al. 2011a. Genetic analyses of atypical *Toxoplasma gondii* strains reveal a fourth clonal lineage in North America. Int J Parasitol. 41(6):645–655.
Khan A, et al. 2011b. A monomorphic haplotype of chromosome Ia is associated with widespread success in clonal and nonclonal populations of *Toxoplasma gondii*. MBio 2(6):e00228–e00211.
Koboldt DC, Larson DE, Wilson RK. 2013. Using VarScan 2 for germline variant calling and somatic mutation detection. Curr Protoc Bioinformatics 44:15.14.1–17.
Kong JT, Grigg ME, Uyetake L, Parmley S, Boothroyd JC. 2003. Serotyping of *Toxoplasma gondii* infections in humans using synthetic peptides. J Infect Dis. 187(9):1484–1495.
Krishna R, et al. 2015. A large-scale proteogenomics study of apicomplexan pathogens-*Toxoplasma gondii* and *Neospora caninum*. Proteomics 15(15):2618–2628.
Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(9):1639–1645.
Laity JH, Lee BM, Wright PE. 2001. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 11(1):39–46.
Lei T, Wang H, Liu J, Nan H, Liu Q. 2014. *ROP18* is a key factor responsible for virulence difference between *Toxoplasma gondii* and *Neospora caninum*. PLoS One 9(6):e99744.
Leineweber M, et al. 2017. First characterization of the *Neospora caninum* dense granule protein GRA9. Biomed Res Int. 2017:1.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.
Lindsay DS, Dubey JP, Duncan RB. 1999. Confirmation that the dog is a definitive host for *Neospora caninum*. Vet Parasitol. 82(4):327–333.
Lorenzi H, et al. 2016. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. Nat Commun. 7:10147.
Lovett JL, Howe DK, Sibley LD. 2000. Molecular characterization of a thrombospondin-related anonymous protein homologue in *Neospora caninum*. Mol Biochem Parasitol. 107(1):33–43.
Ma L, et al. 2017. *Neospora caninum ROP16* play an important role in the pathogenicity by phosphorylating host cell STAT3. Vet Parasitol. 243:135–147.
Ma L, et al. 2017. Rhoptry protein 5 (*ROP5*) is a key virulence factor in *Neospora caninum*. Front Microbiol. 8:370.
Maley SW, et al. 2003. The pathogenesis of neosporosis in pregnant cattle: inoculation at mid-gestation. J Comp Pathol. 129(2-3):186–195.
McInnes LM, Irwin P, Palmer DG, Ryan UM. 2006. In vitro isolation and characterisation of the first canine *Neospora caninum* isolate in Australia. Vet Parasitol. 137(3-4):355–363.
McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
Miller CM, Quinn HE, Windsor PA, Ellis JT. 2002. Characterisation of the first Australian isolate of *Neospora caninum* from cattle. Aust Vet J. 80(10):620–625.
Naguleswaran A, et al. 2001. *Neospora caninum* microneme protein *NcMIC3*: secretion, subcellular localization, and functional involvement in host cell interaction. Infect Immun. 69(10):6483–6494.
Nam HW. 2009. GRA proteins of *Toxoplasma gondii*: maintenance of host-parasite interactions across the parasitophorous vacuolar membrane. Korean J Parasitol. 47(Suppl):S29–S37.
Neer EJ, Schmidt CJ, Nambudripad R, Smith TF. 1994. The ancient regulatory-protein family of WD-repeat proteins. Nature 371(6495):297–300.
Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 7(1):61–80.
O'Rawe J, et al. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5(3):28.
Pabinger S, et al. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 15(2):256–278.

Picoult-Newberg L, et al. 1999. Mining SNPs from EST databases. Genome Res. 9(2):167–174.
Quevillon E, et al. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33(Web Server):W116–W120.
Ramaprasad A, et al. 2015. Comprehensive evaluation of *Toxoplasma gondii* VEG and *Neospora caninum* LIV genomes with tachyzoite stage transcriptome and proteome defines novel transcript features. PLoS One 10(4):e0124473.
Reese ML, Zeiner GM, Saeij JP, Boothroyd JC, Boyle JP. 2011. Polymorphic family of injected pseudokinases is paramount in Toxoplasma virulence. Proc Natl Acad Sci U S A. 108(23):9625–9630.
Regidor-Cerrillo J, et al. 2013. Genetic diversity and geographic population structure of bovine *Neospora caninum* determined by microsatellite genotyping analysis. PLoS One 8(8):e72678.
Regidor-Cerrillo J, Pedraza-Diaz S, Gomez-Bautista M, Ortega-Mora LM. 2006. Multilocus microsatellite analysis reveals extensive genetic diversity in *Neospora caninum*. J Parasitol. 92(3):517–524.
Reichel MP, Alejandra Ayanegui-Alcerreca M, Gondim LF, Ellis JT. 2013. What is the global economic impact of *Neospora caninum* in cattle - the billion dollar question. Int J Parasitol. 43(2):133–142.
Reichel MP, Ellis JT. 2002. Control options for *Neospora caninum* infections in cattle–current state of knowledge. N. Z Vet J. 50(3):86–92.
Reichel MP, Ellis JT, Dubey JP. 2007. Neosporosis and hammondiosis in dogs. J Small Anim Pract. 48(6):308–312.
Reumers J, et al. 2012. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol. 30(1):61–68.
Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. 2008. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol. 4(11):e1000213.
Ribeiro A, et al. 2015. An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. BMC Bioinformatics 16(1):382.
Robinson JT, et al. 2011. Integrative genomics viewer. Nat Biotechnol. 29(1):24–26.
Rojo-Montejo S, et al. 2009. Experimental infection with a low virulence isolate of *Neospora caninum* at 70 days gestation in cattle did not result in foetopathy. Vet Res. 40(5):49.
Ruffolo BB, et al. 2016. Isolation and Genotyping of *Toxoplasma gondii* in seronegative urban rats and presence of antibodies in communicating dogs in Brazil. Rev Inst Med Trop Sao Paulo 58(0):28.
Saeij JP, et al. 2006. Polymorphic secreted kinases are key virulence factors in toxoplasmosis. Science 314(5806):1780–1783.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406–425.
Schock A, Innes EA, Yamane I, Latham SM, Wastling JM. 2001. Genetic and biological diversity among isolates of *Neospora caninum*. Parasitology 123(Pt 1):13–23.
Sievers F, Higgins DG. 2014. Clustal omega. Curr Protoc Bioinformatics 48:3.13.11–16.
Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 26(8):1134–1144.
Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 6:175–182.
Sousa S, et al. 2009. Selection of polymorphic peptides from *GRA6* and *GRA7* sequences of *Toxoplasma gondii* strains to be used in serotyping. Clin Vaccine Immunol. 16(8):1158–1169.
Stucky BJ. 2012. SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms. J Biomol Tech. 23(3):90–93.
Talevich E, Kannan N. 2013. Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. BMC Evol Biol. 13(1):117.
Tang H, Thomas PD. 2016. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics 203(2):635–647.
Taylor S, et al. 2006. A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. Science 314(5806):1776–1780.
Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14(2):178–192.
von Bohl A, et al. 2015. A WD40-repeat protein unique to malaria parasites associates with adhesion protein complexes and is crucial for blood stage progeny. Malar J. 14(1):435.
Weber FH, et al. 2013. On the efficacy and safety of vaccination with live tachyzoites of *Neospora caninum* for prevention of neospora-associated fetal loss in cattle. Clin Vaccine Immunol. 20(1):99–105.
Williams DJ, et al. 2007. Immunization of cattle with live tachyzoites of *Neospora caninum* confers protection against fetal death. Infect Immun. 75(3):1343–1348.
Zhao N, Han JG, Shyu CR, Korkin D. 2014. Determining effects of nonsynonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. PLoS Comput Biol. 10(5):e1003592.

Associate editor: Geoff McFadden

Chapter 3

Contribution of introns to the intraspecies diversity associated with the apicomplexan parasite, *Neospora caninum*

Publication:

Calarco, L. and Ellis, J., 2020. *Contribution of introns to the species diversity associated with the apicomplexan parasite, Neospora caninum*. **Parasitol Res**, 1-15.

Certificate:

I certify that the following chapter is largely my own work, although the contributions of other authors are duly recognised. The contribution of other authors are detailed as follows:

- By providing suggestions on topics to be reviewed
- By suggesting references and works to be reviewed as part of this work
- By proof reading draft manuscripts
- By correcting spelling and grammatical errors in drafts
- By providing suggestions to improve writing style and language
- By providing suggestions to improve layout and structure

Otherwise, the core composition of this work is credited to me.

I hereby certify that the above statements are true and correct:

Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.
Larissa Calarco	Prof John Ellis
PhD Candidate	Co-Author

Date: January 2020

Parasitology Research
https://doi.org/10.1007/s00436-019-06561-x

GENETICS, EVOLUTION, AND PHYLOGENY - ORIGINAL PAPER

Contribution of introns to the species diversity associated with the apicomplexan parasite, *Neospora caninum*

Larissa Calarco[1] · John Ellis[1]

Received: 31 July 2019 / Accepted: 19 November 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Neospora caninum is an intracellular parasite considered a leading cause of bovine reproduction failure worldwide, and a serious neurological disease of canines. Transplacental transmission in intermediate hosts is considered the most efficient means of transmission, which strictly involves asexual reproduction. Nonetheless, extensive genetic diversity has been reported within the species. What is yet to be elucidated are the major drivers of such diversity, and their impact on important parasite phenotypes such as virulence. Instead of protein-encoding sequences, genome and transcriptome data were used to investigate SNPs in introns between two distinct *N. caninum* isolates, with reported differences in pathogenicity. Variant analysis identified 840 and 501 SNPs within intergenic regions and introns, respectively, distinctly concentrated on chromosomes VI and XI, whereas the rest of the genome was monomorphic in comparison. Gene ontologies for SNP-dense intron-containing genes included ATP binding, transmembrane transport, protein kinase activity, and transcription and translation processes. This study shows that variation in non-coding DNA is contributing to *N. caninum* intraspecies genetic diversity, and potentially influencing and contributing to important parasite mechanisms. Finally, we present an assembled and annotated *N. caninum* apicoplast genome and show that this essential organelle is highly conserved between the two isolates, and related Coccidia.

Keywords Transcriptomics · Genomics · Structural variation · Non-coding DNA · Apicoplast · SNPs

[Production Note: This paper is not included in this digital copy due to copyright restrictions.]

View/Download from: Publisher's site

Section Editor: Lihua Xiao

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00436-019-06561-x) contains supplementary material, which is available to authorized users.

✉ Larissa Calarco
Larissa.M.Calarco@student.uts.edu.au

[1] School of Life Sciences, University of Technology Sydney, PO Box 123, Broadway, Ultimo, NSW 2007, Australia

Published online: 03 January 2020

Chapter 4

Annotating the 'hypothetical' in hypothetical proteins: *in-silico* analysis of uncharacterised proteins for the apicomplexan parasite, *Neospora caninum*

Publication:

Calarco, L. and Ellis, J. 2019. *Annotating the 'hypothetical' in hypothetical proteins: in-silico analysis of uncharacterised proteins for the apicomplexan parasite, Neospora caninum*. **Vet Parasitology**, 265, 29-37.

Certificate:

I certify that the following chapter is largely my own work, although the contributions of other authors are duly recognised. The contribution of other authors are detailed as follows:

- By providing suggestions on topics to be reviewed
- By suggesting references and works to be reviewed as part of this work
- By proof reading draft manuscripts
- By correcting spelling and grammatical errors in drafts
- By providing suggestions to improve writing style and language
- By providing suggestions to improve layout and structure

Otherwise, the core composition of this work is credited to me.

I hereby certify that the above statements are true and correct:

Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.
Larissa Calarco	Prof John Ellis
PhD Candidate	Co-Author

Date: January 2020

Veterinary Parasitology 265 (2019) 29–37

Contents lists available at ScienceDirect

Veterinary Parasitology

journal homepage: www.elsevier.com/locate/vetpar

Research paper

Annotating the 'hypothetical' in hypothetical proteins: *In-silico* analysis of uncharacterised proteins for the Apicomplexan parasite, *Neospora caninum*

Larissa Calarco*, John Ellis

School of Life Sciences, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia

ARTICLE INFO

Keywords:
Hypothetical protein
In-silico
Annotation
Virulence factors
Adhesin
Transmembrane protein

ABSTRACT

Neospora caninum is a parasite of veterinary and economic importance, affecting beef and dairy cattle industries globally. While this species has been recognised as a serious cause of disease in cattle and dogs for over 30 years, treatment and control options are still not available. Furthermore, whilst vaccination was identified as the most economic control strategy, vaccine discovery programs require new leads to investigate as vaccines.

The current lack of gene annotation available for *N. caninum*, especially compared to the closely related model organism, *Toxoplasma gondii,* considerably hinders vaccine related research. Moreover, due to the high degree of similarity between the two organisms, a significant amount of gene annotation available for *N. caninum* stems from sequence homology between the species. However, there is a plethora of literature identifying conserved virulence factors between members of the Apicomplexa, which suggests that key players are contributing to successful parasite invasion, motility, and host cell attachment.

In this study, bioinformatic approaches classified 125 uncharacterised proteins within the *N. caninum* genome, as transmembrane proteins with signal peptide sequences. Functional annotation assigned enriched gene ontologies for cell-adhesion, ATP binding, protein serine/threonine phosphatase complex, immune system process, antigen binding, and proteolysis. Additionally, 32 of these proteins were also identified as adhesins, or having adhesin-like properties, which were further characterised through the discovery of domains and gene ontology, to reveal their potential functional significance as virulence factors for *N. caninum*. This study identifies a new, small subset of proteins within *N. caninum*, that may be involved in host-cell interaction, parasite adhesion, and invasion, thereby implicating them as potential targets to exploit in the development of control options against the disease.

[Production Note: This paper is not included in this digital copy due to copyright restrictions.]

View/Download from: UTS OPUS or Publisher's site

* Corresponding author at: University of Technology Sydney, School of Life Sciences, Sydney, Australia.
E-mail addresses: Larissa.Calarco@uts.edu.au, Larissa.M.Calarco@student.uts.edu.au (L. Calarco), John.Ellis@uts.edu.au (J. Ellis).

https://doi.org/10.1016/j.vetpar.2018.11.015
Received 30 July 2018; Received in revised form 30 October 2018; Accepted 24 November 2018
0304-4017/ © 2018 Elsevier B.V. All rights reserved.

Chapter 5

Species diversity and genome evolution of the pathogenic protozoan parasite, *Neospora caninum*

Calarco, L.[1] and Ellis, J.[1]

[1] School of Life Sciences, University of Technology Sydney,
Ultimo, NSW 2007, Australia

Under review:

This manuscript has been submitted to a scientific journal with the intention to publish, and has been formatted based on journal requirements.

Certificate:

I certify that the following chapter is largely my own work, although the contributions of other authors are duly recognised. The contribution of other authors are detailed as follows:

- By providing suggestions on topics to be reviewed
- By suggesting references and works to be reviewed as part of this work
- By proof reading draft manuscripts
- By correcting spelling and grammatical errors in drafts
- By providing suggestions to improve writing style and language
- By providing suggestions to improve layout and structure

Otherwise, the core composition of this work is credited to me.

I hereby certify that the above statements are true and correct:

Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.
Larissa Calarco	Prof John Ellis
PhD Candidate	Co-Author

Date: January 2020

Species diversity and genome evolution of the pathogenic protozoan parasite, *Neospora caninum*

Larissa Calarco[*a] and John Ellis[a]

[a] School of Life Sciences, University of Technology Sydney,

PO Box 123, Broadway,

NSW 2007, Australia

***Author for correspondence**: Larissa Calarco

School of Life Sciences, University of Technology Sydney,

Sydney, Australia,

E: Larissa.M.Calarco@student.uts.edu.au

P: +61 2 9514 4161

ABSTRACT

Neospora caninum is a cyst-forming coccidian parasite of veterinary and economical significance, affecting dairy and beef cattle industries on a global scale. Comparative studies suggest that *N. caninum* consists of a globally dispersed, diverse population of lineages, distinguished by their geographical origin, broad host range, and phenotypic features. While intraspecies diversity, and more specifically pathogenic variability, has been experimentally demonstrated in a myriad of studies, the underlying contributors and sources responsible for such diversity have remained nebulous. However, recent large scale sequence and bioinformatics studies have aided in revealing intrinsic genetic differences distinguishing isolates of this species, that await further characterisation as causative links to virulence and pathogenicity. Furthermore, progress on *N. caninum* research as a non-model organism is hindered by a lack of robust, annotated genomic, transcriptomic, and proteomic data for the species, especially compared to other thoroughly studied Apicomplexa such as *Toxoplasma gondii* and *Plasmodium* species.

This review explores the current body of knowledge on intra-species diversity within *N. caninum*. This includes the contribution of sequence variants in both coding and non-coding regions, the presence of polymorphic hotspots in specific genomic windows, and the identification of non-synonymous mutations. The implications of such diversity on important parasite phenotypes such as pathogenicity and population structure are also discussed. Lastly, the identification of potential virulence factors from both *in-silico* and next generation sequencing studies is examined, offering new insights into potential avenues for future research on neosporosis.

Keywords: Apicomplexa; Genomics; Transcriptomics; Population genetics; Hypothetical protein; Virulence

1. Introduction

Neospora caninum is a ubiquitous cyst-forming Apicomplexan parasite that presents a risk to bovine industries worldwide. Since its initial discovery, neosporosis has been identified as the leading cause of bovine reproduction failure worldwide, as well as a serious neurological disease of canids such as dogs (Dubey et al., 1988a; Dubey and Lindsay, 1996; Reichel and Ellis, 2002; Reichel et al., 2007). It is estimated that neosporosis collectively costs the Australian and New Zealand dairy and beef cattle industries over $110 million per year in losses (Miller et al., 2002; Reichel and Ellis, 2002), and approaching US $546 million in the USA (Reichel et al., 2013). *Neospora caninum* infects a wide host range, the lifecycle of which contains several infectious stages including intracellular tachyzoites and bradyzoites (tissue cysts) found in intermediate hosts, and environmentally impervious oocysts shed by definitive hosts such as dogs (Dubey et al., 2006; Dubey et al., 2007).

While neosporosis as a disease has been recognised for over 30 years, treatment and control options remain inadequate and limited. Of the current strategies available, the balance between effectiveness, economic feasibility, and on-farm applicability is of paramount importance. Reichel and Ellis (2006) compared available options for cost and effectiveness, and concluded that for within-herd prevalences <18% for *N. caninum* infections, inaction was the optimal economic decision over a one year period. Contrastingly for within-herd infection prevalences ≥18%, vaccination was considered the most economic strategy. While vaccination with a killed tachyzoite has been described as highly successful in rodent models (Liddell et al., 1999), it is estimated to be only 50% efficacious in cattle, as well as costly (Romero et al., 2004). However, more successful vaccines, such as those utilising live attenuated tachyzoites, could eliminate the risk of an entire herd aborting, reducing the cost of controlling *N. caninum* to

vaccination alone, and hence offering a more viable option to passivity (Guy et al., 2005; Miller et al., 2005; Williams et al., 2007).

With all things considered, there is clearly a need for and immense value in investigating and elucidating potential vaccine candidates and therapeutic targets against neosporosis, to which molecular biology, next generation sequencing (NGS) data, and bioinformatics has much to offer. This review takes a look at the current body of evidence that lends to our understanding of *N. caninum* virulence, intraspecies diversity, and population genetics, as well as the challenges and limitations associated with conducting and building on such research.

2. Biological and genetic diversity of *N. caninum*

Significant levels of variation within *N. caninum* populations exist, as genotypic and phenotypic features aren't rigorously conserved within the species (Al-Qassab et al., 2010b). Such diversity is expected due to the broad intermediate host range, worldwide distribution, and the capacity for sexual reproduction. While only minor differences may exist in ultrastructure, *N. caninum* isolates appear to vary in their biological and genetic characteristics, of which many studies have reported differences (Atkinson et al., 1999; Barber et al., 1995; Barr et al., 1991; Lindsay and Dubey, 1989; McInnes et al., 2006; Rojo-Montejo et al., 2009b). Vertical transmission, and more specifically endogenous transplacental transmission, in cattle and other ruminants during pregnancy is considered not only the predominant route of transmission, but also essential for parasite maintenance within in a herd (Gonzalez-Warleta et al., 2018; Trees and Williams, 2005; Williams et al., 2009). Furthermore, the importance and frequency of sexual reproduction in definitive hosts to parasite biology, epidemiology, and evolution remain unclear, although the presence of dogs on cattle farms are consistently reported as a risk factor for

bovine neosporosis (Dijkstra et al., 2001; Ribeiro et al., 2019; Robbe et al., 2016). What still requires attention therefore, is a better understanding of the major, underlying contributors responsible for reported genetic diversity, and their relevance to important parasite phenotypes such as virulence and population genetics.

2.1. Distinguishing isolates of *N. caninum* by their biological behaviours

One of the earliest observations suggesting the existence of heterogeneity within N. caninum, was based on in vitro growth rate experiments. Schock et al. (2001) assessed relative growth rates of six N. caninum isolates by measuring ^{3}[H]uracil uptake, and reported significant differences. NC-Liverpool demonstrated significantly faster growth rates in comparison to the other isolates investigated, being twice as fast as NC1, whereas the Swedish bovine isolate, NC-SweB1, was the slowest. A subsequent study demonstrated a lower tachyzoite yield and viability rate of a Spanish isolate of low virulence, NC-Spain 1H, compared to NC1, which also more extensively destroyed the cell monolayer (Rojo-Montejo et al., 2009b).

Initial observations and comparisons in mice and cattle also indicated that isolates of this species exhibit biologically distinct behaviours. NC-Liverpool for example, has been thoroughly studied in murine models, where infection leads to characteristic, clinical neosporosis of weight loss, lethargy and eventual death (Atkinson et al., 1999). Atkinson *et al*. (1999) reported the presence of moderate to extreme necrotic brain lesions in almost all experimental mice infected with NC-Liverpool, as well as hindlimb paralysis and disco-ordinated movement. This was in contrast to mice infected with NC-SweB1 that presented with a milder meningoencephalitis and less severe symptoms. Subsequently, Miller et al. (2002) showed that the NC-Nowra isolate was low in virulence due to the absence of clinical signs of infection in the calf from which it was isolated. Successively,

live NC-Nowra tachyzoites were used to vaccinate mice prior to pregnancy, the result of which reduced transplacental transmission of a challenge strain by 80-90% (Miller et al., 2005). Subsequent inoculation of the tachyzoites into cattle prior to pregnancy conferred very high levels of protection against foetal loss, following a challenge by the NC-Liverpool isolate (Weber et al., 2013; Williams et al., 2007).

Additional studies have investigated fetal mortality rates in cattle infected with *N. caninum*, where for example the NC1 isolate was shown to induce fetal death in heifers (Dubey et al., 1988b; Dubey et al., 1992; Innes et al., 2001; Maley et al., 2003), whereas the low virulence NC-Spain 1H isolate does not (Rojo-Montejo et al., 2009a). However, virulent NC-Spain 7 also results in higher fetal mortality rates and IgG response in infected cattle, compared to cattle inoculated with NC1 (Caspe et al., 2012). Furthermore, a recent study examined early infection dynamics of heifers infected at mid-gestation, with low virulence NC-Spain 1H and high virulence NC-Spain 7 (Jimenez-Pelayo et al., 2019). This study demonstrated that NC-Spain 7 tachyzoites reached the placenta earlier, and caused lesional development and fetal mortality following transmission, whereas in comparison, NC-Spain 1H tachyzoites were delayed in reaching the placenta, and lesion development and transmission was not observed during early infection. Based on the knowledge afforded by such comparative studies, elucidating the existence of underlying genetic variation between *N. caninum* isolates with marked phenotypic differences, presents an invaluable avenue of research warranting investigation.

2.2. Characterising the underlying genetic diversity within the species

A plethora of molecular methods over the last three decades have lent themselves to elucidating potential sources of biological diversity in *N. caninum* at the genetic level. These include randomly amplified polymorphic DNA PCR (RAPD-PCR) (Atkinson et

al., 1999; Davison et al., 1999; Schock et al., 2001; Spencer et al., 2000), multilocus mini- and microsatellite analysis (Basso et al., 2009; Regidor-Cerrillo et al., 2013; Regidor-Cerrillo et al., 2006), and multiplex PCR (Al-Qassab et al., 2010a). Initially, various molecular techniques such as these were employed to study common molecular markers including internal transcribed spacer 1 sequence (ITS1), 18S ribosomal DNA, and the Nc5 repeat. However, little to no nucleotide variation was reported for many of these molecular targets studied using a range of isolates, varying in host species, geographical origin, and pathogenicity.

For example, no significant sequence variation was reported in various studies comparing 18S rDNA from NC-Liverpool, NC1, NC-SweB1, and four other bovine isolates (BPA1, BPA2, BPA3, and BPA4) (Barber et al., 1995; Holmdahl et al., 1997; Marsh et al., 1995; Stenlund et al., 1997). Such results show that this marker is conserved within *N. caninum*, since its evolutionary divergence from its closest ancestors such as *T. gondii* (Marsh et al., 1995). Furthermore, while the repetitive Nc5 sequence and ITS1 region are the most commonly used markers for detection of *N. caninum* through PCR (Dubey and Schares, 2006), only some isolates exhibit minor sequence variation in these regions, and hence are not sufficient enough for differentiation purposes (Al-Qassab et al., 2010b; Gondim et al., 2004; Slapeta et al., 2002). Sequence analysis of the ITS1 region however was useful in establishing that equine isolates of *Neospora hughesi*, the only other known species of the *Neospora* genus, were distinct from both bovine and canine *N. caninum* isolates (Dubey et al., 2001).

Initially, RAPD-PCR was used in various studies to distinguish *N. caninum* from related coccidia, including *T. gondii*, *Sarcocystis* species (Guo and Johnson, 1995; Schock et al., 2001), *Hammondia heydorni* (Sreekumar et al., 2003), *Cryptosporidium parvum* (Schock et al., 2001), and *N. hughesi* (Spencer et al., 2000). This technique was

then exploited to investigate intraspecies genetic diversity within the species, where Atkinson *et al.* (1999) was first able to generate distinct profiles for NC-Liverpool and NC-SweB1. After obtaining the first isolate of *N. caninum* from cattle in the UK, Davison *et al*. (1999) reported differences between this designated NC-LivB1 isolate, and NC-Liverpool, NC1, and NC-SweB using RAPD-PCR techniques. Shortly after, Schock *et al.* (2001) analysed DNA from six *N. caninum* isolates from both bovine and canine hosts, as well as three strains of *T. gondii*, *Sarcocystis* sp., and *C. parvum* using RAPD-PCR techniques. A total of 222 of 434 markers were found to be conserved within the *N. caninum* isolates, however distinct from the other Apicomplexa studied. Furthermore, 54 unique *N. caninum* markers exhibited sufficient variability to distinguish each isolate. The RAPD-PCR results also identified *T. gondii* as *N. caninum's* closest neighbour, but failed to cluster the individual isolates based on either host or geography.

The markers and techniques discussed that were initially employed to explore genetic diversity within *N. caninum*, while of value at the time, in most cases failed to provide a significant level of genetic heterogeneity between isolates (Al-Qassab et al., 2010b; Regidor-Cerrillo et al., 2006). Furthermore, results produced by RAPD-PCR experiments in particular, were to be interpreted cautiously, due to considerations such as DNA purity, primer selection, and extraneous DNA (Al-Qassab et al., 2010b; Gondim et al., 2004; Spencer et al., 2000). The lack of discriminatory power offered by these methods, also meant that epidemiological and intraspecies population structure analyses remained uncharted and unresolved for *N. caninum*.

Subsequently, the gold standard for assessing genetic diversity in *N. caninum*, became the analysis of polymorphic mini- and microsatellites (Al-Qassab et al., 2010b; Regidor-Cerrillo et al., 2006). Such repetitive elements offered a higher degree of discriminatory power compared to other routinely used phylogenetic markers, where

distinctive patterns exhibited by various isolates proved useful in distinguishing and characterising genetic diversity within the species. Mini- and microsatellite techniques have also been extensively used for genotyping other Apicomplexan species. For *T. gondii*, Ajzenberg *et al.* (2010) developed a simple genotyping method that incorporated 15 microsatellite markers across 11 chromosomes in one multiplex PCR assay. These markers were validated for 26 *T. gondii* reference isolates, and were able to differentiate isolates both at the typing level (i.e. the three main clonal lineages compared to atypical isolates), and the fingerprinting level for distinguishing related isolates within lineages.

Regidor-Cerrillo *et al.* (2006) first identified and analysed 12 microsatellites for nine *N. caninum* isolates of different host and geographical origin. Multi-locus microsatellite analysis revealed unique profiles for each isolate, where 11 of the 12 markers were subsequently applied to clinical samples resulting in the detection of new alleles (Pedraza-Diaz et al., 2009). A larger subsequent study investigated 25 cultured *N. caninum* isolates and reported variation within repetitive sequences of eight of the 27 loci studied (Al-Qassab et al., 2009). A multiplex PCR assay was consequently developed incorporating three minisatellite and three microsatellite markers, presenting a simple, efficient, and informative method for genotyping and distinguishing new and existing *N. caninum* isolates (Al-Qassab et al., 2010a).

These genotyping techniques have also contributed to the study of epidemiology and population genetics within the species (Basso et al., 2009; Basso et al., 2010; Regidor-Cerrillo et al., 2013; Regidor-Cerrillo et al., 2006). A large scale genotyping study evaluated nine microsatellite markers for a total of 108 reference and clinical *N. caninum* isolates, collected over a 10 year period from four countries and two continents (Regidor-Cerrillo et al., 2013). A total of 96 microsatellite multilocus genotypes across seven loci revealed extensive levels of genetic diversity across the samples. Data analysis suggested

that these microsatellite markers were partially correlated with geographical origin, with sub-structuring present for each country population of samples. Genetically distinct clustering of microsatellite genotypes was also found to be associated with abortion, and the results of the study suggested a clonal propagation of microsatellite markers for Spanish *N. caninum* isolates in cattle.

Basso *et al*. (2010) characterised DNA extracted from the brains of 18 foetuses infected with *N. caninum* from epidemic abortion storms, using ten microsatellite markers. The same microsatellite pattern was present in all foetuses from each abortion outbreak, and unique to each herd. These results, in conjunction with avidity analysis, provided evidence that infection had originated from a common source, and therefore that horizontal transmission was a determinant of epidemic abortions. Recently, Cabrera *et al.* (2019) reported that four isolated Uruguayan strains of *N. caninum*, represented three distinct genetic lineages, as determined by microsatellite typing. Phylogenetic analysis further revealed that three of these four strains clustered closely with strains from regional Argentina and Brazil. As the remaining Uruguayan strain was found to group with an unrelated cluster, this suggested that the local strains were from multiple origins.

3. New perspectives: Elucidating the degree of sequence variation between *N. caninum* isolates

The popular typing methods used in genetic diversity studies, as discussed in Section 2.2, commonly exploit repetitive elements such as mini- and microsatellites. While valuable, such techniques generally target non-coding sequences, and their influence on parasite biology and phenotype remains largely unknown. What still remains to be addressed is whether a relationship exists between genetic markers, and heterogeneity in clinical manifestations in infected hosts (Al-Qassab et al., 2010b; Goodswen et al., 2013).

Alternatively, investigating variation present within protein-coding genes of a species can expand our understanding of whether such mutations alter gene function, and consequently an organism's phenotype.

Until recently, the presence of polymorphisms, including single nucleotide polymorphisms (SNPs) and insertions and deletions (indels), within coding sequences of *N. caninum* isolates remained unexplored, representing a gap in knowledge that has been extensively studied in related apicomplexan parasites. For example based on genome-wide SNP studies, the within-lineage variation distinguishing three of the four major lineages of *T. gondii* in the Northern Hemisphere, is <0.01%, compared to an estimated 1-3% variation present between lineages (Boyle et al., 2006). Furthermore, extensive research has revealed that only a limited number of sexual recombination events within the feline definitive host of *T. gondii*, are sufficient to influence population structure and the evolution of pathogenic variability between linages (Boyle et al., 2006; Grigg et al., 2001; Khan et al., 2011a; Khan et al., 2009). In the malaria causing *Plasmodium falciparum*, mutations in various molecular markers was identified as part contributors to the growing global threat of drug resistance. For instance, mutations in the propeller domain of the Kelch 13 gene were associated with artemisinin resistance, hence classifying this domain as a useful marker to monitor the emergence of resistance across South East Asia (Ariey et al., 2014).

3.1. The identification and annotation of hypervariable gene hotspots in the *N. caninum* genome

To address the important, and until now, neglected question of the impact of sequence variation in protein-coding genes on *N. caninum* diversity, Calarco *et al.* (2018) generated NGS data from the tachyzoite lifecycle stage to subsequently perform *in-silico* analyses.

RNA-seq data produced for NC-Liverpool and NC-Nowra isolates was used in a variant analysis pipeline (Figure 1), which resulted in the identification of more than 3,000 differentiating SNPs. The allocation of this SNP callset within the *N. caninum* genome revealed their concentration within specific genomic windows on chromosomes VI, XI, and XII, comprising a total of 19 tachyzoite-associated SNP "hotspots" within transcriptionally active coding regions.

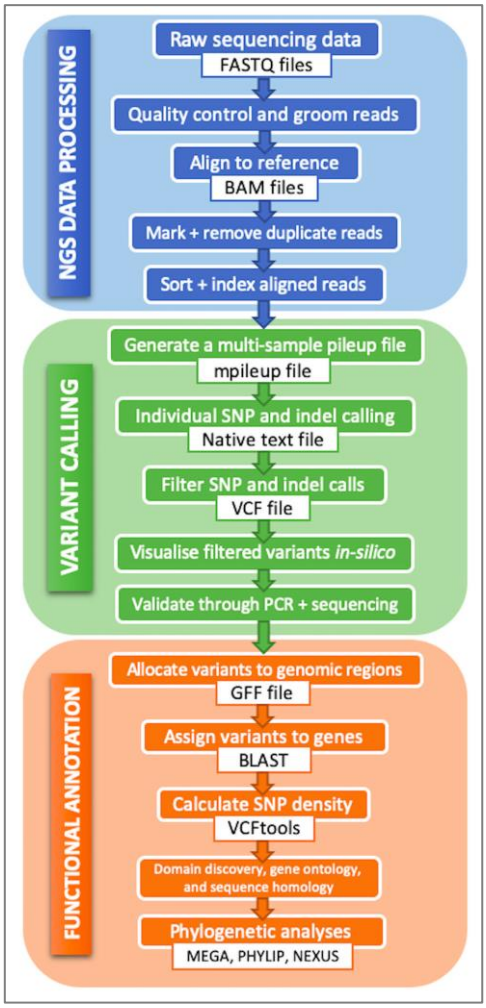

Figure 1. Variant analysis pipeline for detecting SNPs and indels, from raw high-throughput sequencing data generation through to functional analysis of identified variants. Sequence reads produced by next generation sequencing (NGS) platforms, require initial quality control. This can involve trimming the ends of reads where base quality tends to decrease, discarding short sequences, and removing low quality base

calls. Groomed reads can then be aligned to a reference genome or transcriptome, where optimising default alignment parameters can help improve overall mapping quality, especially for non-model organisms such as *N. caninum*. Aligned reads can then be further processed to remove duplicate reads produced through PCR, followed by sorting and indexing BAM files, which are additional steps required by many variant callers. For multi-sample variant calling, an 'mpileup' file can subsequently generated from multiple aligned BAM files, which can then be fed into a variant calling algorithm to detect SNPs and indels, in comparison to a reference FASTA file. Filtered variants should then be visualised to assess their quality, and a subset of variants should be selected for confirmation through PCR and sequencing analysis. Annotating high-quality variant calls can consequently involve allocating them to coding and non-coding genomic regions, performing BLAST analyses to assign polymorphic loci to genes, and also identifying protein sequence domains and repeats, and gene ontologies to assign functional significance to variable loci.

Annotation of these hypervariable genes, many of which contained non-synonymous mutations, revealed their implication in protein-protein interactions, transcription and translation processes, protein binding, ribosomal subunit formation, kinase activity, and protein phosphorylation. Moreover, sequence annotation of the hotspot genes revealed noteworthy protein superfamilies and domains such as WD40 repeat containing (IPR036322), ARM-like helical (IPR011989), P-loop containing nucleoside triphosphate hydrolase (IPR027417), ABC transporter (IPR036640), EF-hand calcium binding domain (IPR018247), AAA+ ATPase domain (IPR003593), and tetratricopeptide repeat (TPR; IPR019734).

The SNP-dense protein-coding genes identified in this study, many of which were characterised as non-synonymous mutations, could represent transcriptionally active, novel determinants of tachyzoite virulence, that distinguish *N. caninum* isolates with reported differences in pathogenicity. Upon further sequence analysis, various hotspot genes also contained transmembrane (TM) domains and/or a signal peptide (SP), indicating the presence of structural sequence features often associated with proteins that are located in the membrane or excreted/secreted that may be potential vaccine candidates.

Interestingly, also reported in this study was almost a complete absence of sequence polymorphisms located within routinely used genotyping markers for *T. gondii*, including surface antigens (SAG1 and SRS2), dense granule proteins (GRA; GRA6), and rhoptry proteins (ROP; ROP30 and ROP37). These results supported previous studies investigating such markers, where Marsh *et al.* (1999) reported complete sequence conservation in genes encoding SAG1 and SRS2, between five *N. caninum* isolates from different continents and host species. Similarly, Walsh *et al.* (2001) compared *GRA6* and *GRA7* gene sequences between four bovine and canine isolates, and found no nucleotide differences.

3.2. The contribution of non-coding SNPs to *N. caninum* intraspecies genetic diversity

The metabolically expensive presence and hence potential functional importance of non-coding DNA has been a topic of growing debate and contention. While initially deemed "junk DNA", non-coding sequences are being elevated in status as potential contributors to the evolutionary diversity and protein repertoire of various species (Fedorova and Fedorov, 2003; Gilbert, 1985; Irimia and Roy, 2008). Furthermore, it was suggested that

sequence polymorphisms within non-coding DNA have the potential to influence genotype-phenotype relationships and gene expression (Cooper, 2010).

Calarco and Ellis (2020) exploited both genomic and transcriptomic data to investigate SNPs located within non-coding and intron regions between isolates. *In-silico* variant analysis was performed using NC-Nowra genomic reads, aligned against the NC-Liverpool reference genome from ToxoDB. The results suggested that the *N. caninum* genome was largely monomorphic, however hypervariable regions were identified on chromosomes VI and XI. Within the variable hotspots, more than half of the identified SNPs on chromosome VI were located within introns, and approximately 70% of SNPs on chromosome XI were in either introns or non-coding genomic regions. The annotation of these hypervariable genes containing SNP-dense introns, revealed gene ontologies (GOs) related to transcription and translation processes, protein kinase activity, ATP binding and ATPase activity, transmembrane transport, and ion binding. There was also a distinct overlap of and correlation between genes containing SNP-dense introns and coding regions. Prioritisation based on SNP density within their respective introns, resulted in the identification of a set of functionally significant genes located on chromosomes VI and XI, including an ABC transporter involved in transmembrane transport and ATP binding (*NCLIV_015830*), a calcium-activated ion channel protein with eight TM domains and a SP (*NCLIV_015820*), and a kinesin-like protein involved in microtubule movement (*NCLIV_056770*).

Further to the topical discussion on non-coding DNA, the role of introns in evolutionary diversity through mechanisms of alternative splicing (AS) is gaining momentum and attracting debate. This includes the impact of phenomena such as exon shuffling, intron retention, mRNA surveillance, transcriptional regulation of gene expression, and recombination (Duret, 2001; Fedorova and Fedorov, 2003; Gilbert, 1985;

Lynch and Richardson, 2002). With respect to protozoan parasites, approximately 22% of protein-coding genes in *T. gondii* (Yeoh et al., 2015), and 16% in *Plasmodium* species (Iriko et al., 2009) are estimated to undergo AS, where studies suggest that AS mechanisms can generate distinct, functionally significant transcripts in Apicomplexan parasites, where such gene products are also considered as prospective vaccine candidates (Agarwal et al., 2011; Delbac et al., 2001; Gabriel et al., 2015; Kern et al., 2014).

By implementing *in-silico* structural variant identification pipelines, Calarco and Ellis (2020) reported 315 and 503 deletions by the Pindel and BreakDancer algorithms respectively, spanning more than 50bp between RNA-seq reads generated from NC-Liverpool and NC-Nowra tachyzoites. Many of these predicted deletions were subsequently identified as introns, present in transcripts from NC-Liverpool, however absent from, and likely removed through splicing in, NC-Nowra transcripts. The corresponding genes containing these introns, presented GOs for RNA processing, protein binding, proteolysis, methyltransferase activity, and oxidation-reduction process, where many also contained TM domains and/or a SP.

In addition, many predicted *in-silico* deletions were instead in regions either just before or after a gene sequence (i.e. non-coding or intergenic sequences). Annotations for such genes including those putatively encoding microneme proteins (MIC3, NCLIV_010600; MIC8, NCLIV_062770), a protein kinase (NCLIV_050650), and a translation initiation factor subunit protein (NCLIV_011760). It was suggested that these deletions may belong to regulatory elements that influence gene expression, or alternatively that reference gene sequence annotations are inaccurate. With respect to functionally important micronemes, following the discovery of the *N. caninum* MIC8 protein, which shows a high similarity to that of its homologue in *T. gondii*, this protein was reported to play a critical role in host-cell invasion, and also to form a complex with

NcMIC3, most likely during transportation (Wang et al., 2017). Subsequently, Zhang *et al.* (2019) demonstrated that immunising mice with recombinant NcMIC3 and NcMIC8, provided at least part protection against neosporosis following a challenge with *N. caninum* tachyzoites, with mice also exhibiting a lower parasite burden in brain tissue. Such studies demonstrate the importance of microneme proteins in not only host cell invasion and interaction, but also as vaccine candidates against neosporosis.

Overall, these studies revealed and characterised the contribution of variable, non-coding and intron DNA, to the intraspecies diversity reported for *N. caninum*. Such research improves upon our understanding of the underlying genetic forces occurring in this species. It also suggests that mechanisms of alternative splicing, specifically intron retention in transcripts, distinguish isolates of this species, and may play a larger role in the context of parasite biology, than previously thought. By collating these new studies, we start to gain a clearer picture of the underlying, intrinsic genetic differences occurring in isolates of *N. caninum*.

It is evident that regions on chromosomes VI and XI are highly polymorphic between isolates, in both transcriptionally active tachyzoite-associated genes, as well as intergenic and intron sequences. These data hence expand upon our understanding of the driving forces responsible for important parasite phenotypes and biological mechanisms, conducive to the species' success. In a historical context for example, do these genetic diversity studies indicate that sexual recombination in definitive hosts plays a more important role that previously thought? Do the concentrated polymorphic regions identified point towards sex being of greater significance and influence in parasite evolution, epidemiology, and intraspecies diversity?

4. The evolution and heterogeneity of coccidian genomes

How can we contextualise variation at the genetic and transcriptomic levels to establish its relationship with the population structure, epidemiology, and phylogeny of a species? Revealing the existence of a species' underlying population structure can offer insight into the spread and determination of virulence factors, pathogenicity, and genotyping markers (Khan et al., 2011b). Recent efforts towards identifying genome-wide SNP markers for a range of parasite species, is providing a new framework based on population genetics, with implications in molecular evolution, species divergence, and association studies.

For example, initial genotyping studies for *T. gondii*, and their interpretation in terms of a clonal population structure for the species, were limited by the small number of isolates studied and their host and geographical origins (Darde et al., 1988; Howe and Sibley, 1995; Sibley and Boothroyd, 1992). However, subsequent studies incorporating multilocus techniques for a wider range of isolates, revealed the existence of a more genetically diverse species with a complex population structure (Ajzenberg et al., 2004). Large-scale sequencing efforts for *T. gondii* resulted in over 250 SNPs within random fragment length polymorphism (RFLP) markers, being observed between the three main clonal lineages (Khan et al., 2005). Based on genome-wide SNP studies, it was estimated that the polymorphism rate between the lineages is 0.65% (Boyle et al., 2006), which equates to approximately one variant per 100bp throughout almost all chromosomes (Khan et al., 2005). The three predominate lineages are also classified as either type I, II, or III SNPs (Boyle et al., 2006), with each exhibiting extensive bi-allelism (Grigg et al., 2001; Khan et al., 2005). The distribution of these SNP types also suggests that types I and III represent first and second generation offspring, resulting from a cross between a type II strain and an ancestral strain (Boyle et al., 2006).

It is well understood that diversification, duplication, and expansion of loci is ubiquitous and prevalent in apicomplexan genomes, where this is especially true of protein-coding genes found in a parasite's secretory pathway, and/or expressed on their surface (Blank and Boyle, 2018). The SRS superfamily of surface antigens for example, has been identified as one of the most divergent and rapidly evolving protein families within this phylum (Jung et al., 2004; Manger et al., 1998; Reid et al., 2012; Wasmuth et al., 2009; Wasmuth et al., 2012). As a result, gene duplication and expansion events have been used to study the phenotypic differences and phylogenetic relationships between Apicomplexan species, and more specifically, within the *Toxoplasmatinae* subfamily, with implications in species divergence and evolution (Adomako-Ankomah et al., 2014; Lorenzi et al., 2016; Reid et al., 2012). The genome and transcriptome sequencing of related Coccidia such as *T. gondii*, *N. caninum*, and *H. hammondi* has revealed a high degree of conservation and genomic synteny between these species, with respect to size, protein-coding genes, GC content, and gene catalogue (Reid et al., 2012; Walzer et al., 2013). However, the study of large structural variations at expanded loci have revealed sources of evolutionary divergence within this subfamily.

For example, more than double the number of *SRS* genes were reported for *N. caninum* by Reid *et al*. (2012) compared to *T. gondii*, where there was also a divergence in secreted virulence factors between the species, namely rhoptry kinases. These results suggested that a small set of genes implicated in host-parasite interactions, have influenced the ecological niches and pathogenic capabilities of these species. By comparing the genomes of 62 globally dispersed *T. gondii* isolates to *N. caninum*, *H. hammondi*, and *S. neurona*, Lorenzi *et al.* (2016) reported that these closely related, but phenotypically diverse, parasites could be distinguished based on the tandem amplification and diversification of secreted, pathogenic determinants. Duplicated genes

and copy number variation (CNV) events were specifically in *T. gondii* genes encoding ROPs, MICs, and SAG/SRS, many of which also showed evidence of positive selection based on non-synonymous mutation rates. Comparing the orthologous *GRA*, *ROP*, and *SRS* genes between related Coccidia, suggested that divergence within these genes may be responsible for phenotypic differences observed between these species, whereas MIC genes were highly conserved. Furthermore, Marsh *et al.* (1999) indicated that while the SAG1 and SRS2 protein sequences were conserved across six *N. caninum* isolates originating from different hosts and geographical regions, there was a 6% and 9% difference in these amino acid sequences respectively, when compared to *N. hughesi*. The variation within these protein-coding genes therefore represented additional molecular markers for the distinction of these two *Neospora* species.

4.1 The influence of sequence diversity on *N. caninum* population genetics

Through the *in-silico* identification and laboratory confirmation of a set of SNPs using NGS data, Calarco *et al.* (2018) developed a multilocus sequencing approach for nine *N. caninum* isolates, differentiated by geographical origin, host, and/or reported pathogenicity. This data elucidated a population structure consisting of two major clades, one of which included the virulent NC-Liverpool isolate, and the other containing the low virulence NC-Nowra isolate. Interestingly, the clustering of the isolates based on polymorphic loci did not appear to conform to a pattern based on their geographical segregation or host. Furthermore, the inclusion of *N. hughesi* in the multilocus sequence typing, bolstered support for the two-clade structure for *N. caninum*.

However, there are a limited number of studies documenting and comparing the pathogenicity of the *N. caninum* isolates investigated in this study, in both murine and bovine host models. While this limits the assumptions that can be drawn from such

results, it does highlight the value of SNP data in conjunction with population structure analyses, in enriching our understanding of intraspecies genetic diversity and important parasite phenotypes. These results emphasised that the two representative populations of *N. caninum* presented not only differ in their pathogenicity as previously demonstrated, but are also genetically distinct. Such knowledge can expedite the identification of novel virulence markers, and hence provide candidates for vaccine and drug development against neosporosis.

4.2 Exploring non-nuclear DNA in apicomplexan parasites

Another potential source of genetic variation recently investigated and reported for *N. caninum* was non-nuclear DNA, where Calarco and Ellis (2020) assembled and annotated the apicoplast genome for NC-Liverpool, based on NGS data. The apicoplast is a secondary plastid organelle unique to most species within the apicomplexan phylum, and while it lacks photosynthetic ability, it is essential for parasite survival (Gardner et al., 1991; Howe, 1992; Williamson et al., 1994; Wilson et al., 1996). The published apicoplast genomes of *P. falciparum* (Wilson et al., 1996), *T. gondii* (EMBL accession number U87145), and *Eimeria tenella* (Cai et al., 2003) have been shown to be highly similar with respect to gene catalogue and structure. This includes an inverted repeat region containing duplicated small and large subunit ribosomal RNAs (*SSU rRNA* and *LSU rRNA*), and tRNA genes. The conservation of this organellar genome between genera also extends to the transcription of genes, where half of the apicoplast circle is transcribed in a clockwise direction, and the other half is transcribed counter-clockwise.

While most orthologous genes from these three coccidian species align to one another, with the exception of hypothetical open reading frames (ORFs), their sequences are divergent. Contrastingly, the apicoplast genomes of piroplasmids such as *Theileria*

parva (Gardner et al., 2005) and *Babesia bovus* (Brayton et al., 2007) do not contain an inverted repeat region, and most of their genes are single-copy and transcribed uni-directionally. The sequencing of apicoplast genomes from various apicomplexan species is conducive to elucidating the evolution, phylogeny, population structure, and biological mechanisms of these parasites, which can subsequently contribute to our understanding of the epidemiology and pathogenicity within and between species.

For example, to address limitations associated with nuclear SNP barcodes, including a lack of geographic specificity, Preston *et al.* (2014) exploited the extra-nuclear mitochondrial and apicoplast organelles for *P. falciparum* isolates. Sequencing data from a total of 711 isolates across 14 countries was used to identify high quality SNPs within these genomes, and subsequently analyse the geographic diversity and extent of recombination within the population. A total of 151 mitochondrial and 488 apicoplast SNPs were documented, where variation within these combined non-nuclear genomes resulted in a 23-SNP barcode able to discriminate between the region of sample origin with an accuracy of 92%. High linkage disequilibrium between the catalogued SNPs supported the co-transmitted and non-recombining nature of the organelles, revealing novel haplotypes unique for different geographic regions.

Variant analysis of reads from NC-Nowra aligned to an assembled NC-Liverpool apicoplast genome, revealed that this organelle is highly conserved between these isolates with marked biological differences. Calarco and Ellis (2020) only reported three SNPs and one insertion three base pairs long present in NC-Nowra reads across the ~35kb apicoplast genome. The SNPs were located within RNA polymerase β subunit (*rpoB*), RNA polymerase β' subunit (*rpoC2.2*), and ORF-F, and the insertion was positioned just before the start codon of the *tufA* gene. High sequence similarity was also reported between the NC-Liverpool apicoplast genes when aligned to closely related coccidia such

as *T. gondii*, *H. hammondi*, *S. neurona*, and *Cryptosporidium suis*, with identities between 79-93%. The structure, gene content, and arrangement of the NC-Liverpool apicoplast genome was also highly similar between these related coccidia and *Plasmodium* species, suggesting the evolutionary conservation of this essential organelle. Figure 2 presents the NC-Liverpool *N. caninum* annotated apicoplast genome published by Calarco and Ellis (2020) (MK770339), as aligned to a reference apicoplast genome sequence from *T. gondii* (U87145), which highlights how the nucleotide sequence, and gene arrangement and catalogue of this organelle are highly conserved between these related species.

Figure 2. Comparison of the gene catalogue and structure of the *N. caninum* and *T. gondii* apicoplast genomes. The NC-Liverpool apicoplast genome (accession MK770339) was aligned to a *T. gondii* reference apicoplast genome (accession U87145) in Geneious Prime (www.geneious.com; version 2020.0.4). Genes are presented in yellow, rRNAs are in red, and tRNAs are shown in pink. The top consensus track shows the percentage identity between the two apicoplast nucleotide sequences, where the green represents 100% sequence similarity, yellow is between 30-100% sequence similarity, and red is <30% sequence similarity at each base. This figure highlights how the ~35 kb apicoplast genome is highly conserved between these two closely related species.

5. *Neospora caninum* as a non-model organism: What are we still missing?

One of the greatest challenges currently plaguing the advancement of treatment and control strategies against neosporosis, is at least partly attributable to the status of *N. caninum* as a non-model organism. Consequently, there are limited resources available for the species, and for the data that has been generated, there is the question of the robustness, accuracy, and completeness of such records. This is further compounded by the lack of gene and protein annotations presented in reference resources and databases, where putative gene descriptions commonly stem from sequence homology with closely related and thoroughly studied Apicomplexa. Currently, more than 60% of genes predicted to comprise the *N. caninum* genome remain uncharacterised, and instead are described as coding for 'hypothetical' or 'unnamed' proteins (Calarco and Ellis, 2019).

The SNP hotspot regions identified by Calarco *et al.* (2018) mostly contained genes that were simply labelled "hypothetical" or "unspecified". Annotation using several *in-silico* tools, led to gene classification by homology, protein families, domains, and/or GO terms. However, one quarter of the SNP hotspot genes still remained unannotated after all methods were exhausted. This clearly highlights how such genetic diversity analyses are only as valuable as the annotation available for important loci identified, and hence overall for the organism under investigation. It also alludes to the disturbing and sizeable existence of uncharacterised, theoretically significant proteins that await description for *N. caninum*.

5.1. Virulence factors of Apicomplexa

Members of the Apicomplexa, whilst biologically specialized, share a plethora of cellular and molecular characteristics that are conducive to their success as intracellular parasites.

This includes chiefly an apical complex, and molecules that aid in parasite motility, host cell adhesion, and subsequent invasion (Huynh et al., 2014). Rhoptries, micronemes, and dense granules are the three secretory vesicles constituting the apical organelle, which release effector proteins essential for host cell recognition, invasion, and modification, and are therefore instrumental in the success of parasite virulence (Blackman and Bannister, 2001).

Rhoptry proteins (ROPs) are amongst the most comprehensively studied family of proteins in Apicomplexa, where many have been identified as key virulence factors (Talevich and Kannan, 2013). These include ROP18, which modulates parasite growth, and ROP5, which together play a role in impeding the host immune response (Behnke et al., 2012; Saeij et al., 2006; Talevich and Kannan, 2013). The microneme (MICs) family of proteins can be classified by the presence of TM domains, a SP, and adhesive motifs. Such structural features within these secreted proteins enable the parasite to interact with host cell receptors, thereby facilitating successful attachment to and invasion of host cells. In *T. gondii*, the MIC2 complex has been implicated in gliding motility and cell invasion (Jewett and Sibley, 2003), and has hence been acknowledged as a major virulence determinant in *Toxoplasma* infection, where MIC2-deficient parasites could be exploited in the development of a live attenuated vaccine. Furthermore, while MIC2 was initially described by Lovett *et al.* (2000), it remains listed as a "hypothetical protein' in NCBI, UniProt, and ToxoDB reference databases. Following host cell invasion, dense granule (GRA) proteins are released and serve to modify the conditions of the parasitophorous vacuole membrane and environment, ensuring the successful intracellular maintenance and replication of the parasite (Cesbron-Delauw, 1994; Mercier et al., 2005; Mercier et al., 2002). A total of 12 GRA proteins have been identified in *T. gondii*, many of which contain a SP sequence between 25-30 amino acids, and/or TM domains, and as such are

classified as excretory/secretory antigens, associated with the secretory pathway of the parasite (Ahn et al., 2005; Cesbron-Delauw, 1994; Mercier et al., 2005; Michelin et al., 2009; Rome et al., 2008).

Many key players contributing to crucial parasite mechanisms of invasion, motility, and adhesion, contain distinct sequence features that can be identified *in-silico*. For example, proteins containing TM domains are implicated in a range of crucial biological processes such as cell attachment, invasion, and molecule transport (Reynolds et al., 2008), whereas proteins transported to secretory organelles commonly require an N-terminal SP (Chen et al., 2008). Such aforementioned effector molecules implicated in parasite invasion, host-cell signaling modulation, and adhesion, generally contain important structural features such as these, and have hence been identified as potential vaccine candidates for apicomplexan parasites (Kim and Weiss, 2004). In terms of gene expansion and diversification in *T. gondii*, and its relevance to phenotypic diversity within and between species, Adomako-Ankomah *et al.* (2014) reported that a total of 42 of 53 tandemly expanded loci identified were predicted to be part of the parasite's secretory pathway, and 29 of these contained an N-terminal SP. Consequently, how can this body of knowledge facilitate the detection of crucial virulence factors buried within the *N. caninum* proteome, that still await identification?

5.2. The value of bioinformatics in the quest to resolve the 'hypothetical'

The advent of second and third generation sequencing technologies has seen the mass generation and availability of data for a wide range of organisms, increasingly so for *N. caninum*. To accommodate for this burgeoning field of technology, *in-silico* tools and pipelines are constantly being developed and optimised, to ensure researchers are able to process, analyse, and make such data publicly available. Many well studied and non-model organisms have benefitted from *in-silico* analysis of hypothetical proteins. For

example, Oladele *et al.* (2011) identified several sequences that could represent biomarkers of malaria, through bioinformatics analysis of hypothetical *P. falciparum* proteins. Syn *et al.* (2018) presented an *in-silico* pipeline to identify *T. gondii* proteins influencing host cell epigenetic regulation. The tools used aimed to identify proteins secreted via classical and non-classical pathways, predict proteins localised to the nucleus, and prioritise proteins associated with epigenetic regulation. From a total of 8313 proteins in the *T. gondii* proteome, 57 proteins were predicted to be involved in both secretory pathways, and epigenetic functions.

By combining current knowledge of conserved apicomplexan virulence factors and their sequence features, with suitable bioinformatics tools to facilitate their identification, Calarco *et al*. (2019) curated a list of proteins described as "hypothetical" or "unnamed" from reference resources. A total of 4,264 uncharacterised proteins were then subjected to a range of tools for classification by topology, sequence homology, adhesin-like properties, GOs, and protein domains. From this initial list, 125 uncharacterised *N. caninum* proteins were predicted to contain TM domains and a SP sequence. After these protein sequences were submitted to a malarial adhesins predictor, a final set of 32 proteins were classified as adhesin-like TM proteins with a SP sequence. Following functional annotation of prioritised proteins, enriched GOs included cell adhesion, proteolysis, antigen binding, protein serine/threonine phosphatase complex, locomotion, and ATP binding. Various hypothetical proteins annotated through this workflow contained structural features known to be implicated in important parasite mechanisms, such as those described as MICs, GRAs, proteases, peptidases, and surface antigens.

The development of a bioinformatics workflow in this study, resulted in the prioritisation of all predicted, uncharacterised proteins within the *N. caninum* proteome,

based on sequence structure and features of known, conserved apicomplexan virulence factors. This resulted in the identification of a novel and focused subset of proteins, that are potentially involved in biological processes pertaining to parasite motility, adhesion, invasion, signalling, and interaction with host cells. Consequently, these proteins could be targeted in new avenues of investigation in the near future, for the development of treatment and control options against neosporosis. This includes comparing these proteins between *N. caninum* isolates with reported phenotypic differences, at both a molecular and biological level. This includes nucleotide and amino acid sequence diversity through variant detection, gene expression studies, gene knockout experiments, and immunoassays.

6. Summary

Extensive efforts have recently been made to reveal the key contributors to and major sources of, genetic diversity characterising isolates of the pathogenic protozoa, *N. caninum*. All avenues investigated thus far have provided evidence that *N. caninum* exists as a diverse, globally distributed population of isolates, that exhibit genetic and biological heterogeneity. Initial efforts dedicated to elucidating the genetic diversity between *N. caninum* isolates, consisted of analysing repetitive elements and popular phylogenetic molecular markers. However, these studies generally focused on non-coding genomic regions and hence failed to explain the relationship of such markers to important parasite phenotypes. As a result, recent new efforts have been dedicated to the identification of sequence polymorphisms within protein-coding genes, non-nuclear DNA, and introns, to reveal their contribution to and impact on the biological diversity reported between isolates.

Bioinformatics and *in-silico* analysis of sequencing data both original and publicly accessible, has the potential to reveal and contribute to our understanding of apicomplexan virulence factors and possible vaccine and drug targets. It is clear however, that the development of treatment and control options against neosporosis can only progress as fast as the improvement of gene and protein descriptions. Current analyses of the *N. caninum* genome, proteome, and transcriptome are only as valuable as the annotations available for such data, and hence efforts in the near future should be dedicated and prioritised as such.

Acknowledgements

This study was completed by LC in partial fulfillment of the Ph.D. degree at the University of Technology Sydney

Declaration of interest

None

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Adomako-Ankomah, Y., Wier, G.M., Borges, A.L., Wand, H.E., Boyle, J.P., 2014. Differential locus expansion distinguishes Toxoplasmatinae species and closely related strains of Toxoplasma gondii. MBio 5, e01003-01013.

Agarwal, S., Kern, S., Halbert, J., Przyborski, J.M., Baumeister, S., Dandekar, T., Doerig, C., Pradel, G., 2011. Two nucleus-localized CDK-like kinases with crucial roles for malaria parasite erythrocytic replication are involved in phosphorylation of splicing factor. J Cell Biochem 112, 1295-1310.

Ahn, H.J., Kim, S., Nam, H.W., 2005. Host cell binding of GRA10, a novel, constitutively secreted dense granular protein from Toxoplasma gondii. Biochem Biophys Res Commun 331, 614-620.

Ajzenberg, D., Banuls, A.L., Su, C., Dumetre, A., Demar, M., Carme, B., Darde, M.L., 2004. Genetic diversity, clonality and sexuality in Toxoplasma gondii. International journal for parasitology 34, 1185-1196.

Ajzenberg, D., Collinet, F., Mercier, A., Vignoles, P., Darde, M.L., 2010. Genotyping of Toxoplasma gondii isolates with 15 microsatellite markers in a single multiplex PCR assay. J Clin Microbiol 48, 4641-4645.

Al-Qassab, S., Reichel, M.P., Ellis, J., 2010a. A second generation multiplex PCR for typing strains of Neospora caninum using six DNA targets. Mol Cell Probes 24, 20-26.

Al-Qassab, S., Reichel, M.P., Ivens, A., Ellis, J.T., 2009. Genetic diversity amongst isolates of Neospora caninum, and the development of a multiplex assay for the detection of distinct strains. Mol Cell Probes 23, 132-139.

Al-Qassab, S.E., Reichel, M.P., Ellis, J., 2010b. On the biological and genetic diversity in Neospora caninum. Diversity, 2, 27.

Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., Lim, P., Leang, R., Duong, S., Sreng, S., Suon, S., Chuor, C.M., Bout, D.M., Menard, S., Rogers, W.O., Genton, B., Fandeur, T., Miotto, O., Ringwald, P., Le Bras, J., Berry, A., Barale, J.C., Fairhurst, R.M., Benoit-Vical, F., Mercereau-Puijalon, O., Menard, D., 2014. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature 505, 50-55.

Atkinson, R., Harper, P.A., Ryce, C., Morrison, D.A., Ellis, J.T., 1999. Comparison of the biological characteristics of two isolates of Neospora caninum. Parasitology 118 (Pt 4), 363-370.

Barber, J.S., Holmdahl, O.J., Owen, M.R., Guy, F., Uggla, A., Trees, A.J., 1995. Characterization of the first European isolate of Neospora caninum (Dubey, Carpenter, Speer, Topper and Uggla). Parasitology 111 (Pt 5), 563-568.

Barr, B.C., Anderson, M.L., Dubey, J.P., Conrad, P.A., 1991. Neospora-like protozoal infections associated with bovine abortions. Veterinary pathology 28, 110-116.

Basso, W., Schares, S., Barwald, A., Herrmann, D.C., Conraths, F.J., Pantchev, N.,

Vrhovec, M.G., Schares, G., 2009. Molecular comparison of Neospora caninum oocyst isolates from naturally infected dogs with cell culture-derived tachyzoites of the same isolates using nested polymerase chain reaction to amplify microsatellite markers. Veterinary parasitology 160, 43-50.

Basso, W., Schares, S., Minke, L., Barwald, A., Maksimov, A., Peters, M., Schulze, C., Muller, M., Conraths, F.J., Schares, G., 2010. Microsatellite typing and avidity analysis suggest a common source of infection in herds with epidemic Neospora caninum-associated bovine abortion. Veterinary parasitology 173, 24-31.

Behnke, M.S., Fentress, S.J., Mashayekhi, M., Li, L.X., Taylor, G.A., Sibley, L.D., 2012. The polymorphic pseudokinase ROP5 controls virulence in Toxoplasma gondii by regulating the active kinase ROP18. PLoS pathogens 8, e1002992.

Blackman, M.J., Bannister, L.H., 2001. Apical organelles of Apicomplexa: biology and isolation by subcellular fractionation. Molecular and biochemical parasitology 117, 11-25.

Blank, M.L., Boyle, J.P., 2018. Effector variation at tandem gene arrays in tissue-dwelling coccidia: who needs antigenic variation anyway? Curr Opin Microbiol 46, 86-92.

Boyle, J.P., Rajasekar, B., Saeij, J.P.J., Ajioka, J.W., Berriman, M., Paulsen, I., Roos, D.S., Sibley, L.D., White, M.W., Boothroyd, J.C., 2006. Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like Toxoplasma gondii. Proc Natl Acad Sci U S A 103, 10514-10519.

Brayton, K.A., Lau, A.O., Herndon, D.R., Hannick, L., Kappmeyer, L.S., Berens, S.J., Bidwell, S.L., Brown, W.C., Crabtree, J., Fadrosh, D., Feldblum, T., Forberger, H.A., Haas, B.J., Howell, J.M., Khouri, H., Koo, H., Mann, D.J., Norimine, J., Paulsen, I.T., Radune, D., Ren, Q., Smith, R.K., Jr., Suarez, C.E., White, O., Wortman, J.R., Knowles, D.P., Jr., McElwain, T.F., Nene, V.M., 2007. Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog 3, 1401-1413.

Cabrera, A., Fresia, P., Berna, L., Silveira, C., Macias-Rioseco, M., Arevalo, A.P., Crispo, M., Pritsch, O., Riet-Correa, F., Giannitti, F., Francia, M.E., Robello, C., 2019. Isolation and molecular characterization of four novel Neospora caninum strains. Parasitol Res 118, 3535-3542.

Cai, X., Fuller, A.L., McDougald, L.R., Zhu, G., 2003. Apicoplast genome of the coccidian Eimeria tenella. Gene 321, 39-46.

Calarco, L., Barratt, J., Ellis, J., 2018. Genome Wide Identification of Mutational Hotspots in the Apicomplexan Parasite Neospora caninum and the Implications for Virulence. Genome Biol Evol 10, 2417-2431.

Calarco, L., Ellis, J., 2019. Annotating the 'hypothetical' in hypothetical proteins: In-silico analysis of uncharacterised proteins for the Apicomplexan parasite, Neospora caninum. Veterinary parasitology 265, 29-37.

Calarco, L., Ellis, J., 2020. Contribution of introns to the species diversity associated with the apicomplexan parasite, *Neospora caninum*. Parasitol Res, 1-15. http://doi.org/10.1007/s00436-019-06561-x

Caspe, S.G., Moore, D.P., Leunda, M.R., Cano, D.B., Lischinsky, L., Regidor-Cerrillo, J., Alvarez-Garcia, G., Echaide, I.G., Bacigalupe, D., Ortega Mora, L.M., Odeon, A.C., Campero, C.M., 2012. The Neospora caninum-Spain 7 isolate induces placental damage, fetal death and abortion in cattle when inoculated in early gestation. Veterinary parasitology 189, 171-181.

Cesbron-Delauw, M.F., 1994. Dense-granule organelles of Toxoplasma gondii: their role in the host-parasite relationship. Parasitol Today 10, 293-296.

Chen, Z., Harb, O.S., Roos, D.S., 2008. In silico identification of specialized secretory-organelle proteins in apicomplexan parasites and in vivo validation in Toxoplasma gondii. PLoS One 3, e3611.

Cooper, D.N., 2010. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum Genomics 4, 284-288.

Darde, M.L., Bouteille, B., Pestre-Alexandre, M., 1988. Isoenzymic characterization of seven strains of Toxoplasma gondii by isoelectrofocusing in polyacrylamide gels. Am J Trop Med Hyg 39, 551-558.

Davison, H.C., Guy, F., Trees, A.J., Ryce, C., Ellis, J.T., Otter, A., Jeffrey, M., Simpson, V.R., Holt, J.J., 1999. In vitro isolation of Neospora caninum from a stillborn calf in the UK. Res Vet Sci 67, 103-105.

Delbac, F., Sanger, A., Neuhaus, E.M., Stratmann, R., Ajioka, J.W., Toursel, C., Herm-Gotz, A., Tomavo, S., Soldati, T., Soldati, D., 2001. Toxoplasma gondii myosins B/C: one gene, two tails, two localizations, and a role in parasite division. J Cell Biol 155, 613-623.

Dijkstra, T., Eysker, M., Schares, G., Conraths, F.J., Wouda, W., Barkema, H.W., 2001. Dogs shed Neospora caninum oocysts after ingestion of naturally infected bovine placenta but not after ingestion of colostrum spiked with Neospora caninum tachyzoites. International journal for parasitology 31, 747-752.

Dubey, J.P., Buxton, D., Wouda, W., 2006. Pathogenesis of bovine neosporosis. J Comp Pathol 134, 267-289.

Dubey, J.P., Carpenter, J.L., Speer, C.A., Topper, M.J., Uggla, A., 1988a. Newly recognized fatal protozoan disease of dogs. Journal of the American Veterinary Medical Association 192, 1269-1285.

Dubey, J.P., Hattel, A.L., Lindsay, D.S., Topper, M.J., 1988b. Neonatal Neospora caninum infection in dogs: isolation of the causative agent and experimental transmission. Journal of the American Veterinary Medical Association 193, 1259-1263.

Dubey, J.P., Liddell, S., Mattson, D., Speert, C.A., Howe, D.K., Jenkins, M.C., 2001. Characterization of the Oregon isolate of Neospora hughesi from a horse. J Parasitol 87, 345-353.

Dubey, J.P., Lindsay, D.S., 1996. A review of Neospora caninum and neosporosis. Veterinary parasitology 67, 1-59.

Dubey, J.P., Lindsay, D.S., Anderson, M.L., Davis, S.W., Shen, S.K., 1992. Induced transplacental transmission of Neospora caninum in cattle. Journal of the American Veterinary Medical Association 201, 709-713.

Dubey, J.P., Schares, G., 2006. Diagnosis of bovine neosporosis. Veterinary parasitology 140, 1-34.
Dubey, J.P., Schares, G., Ortega-Mora, L.M., 2007. Epidemiology and control of neosporosis and Neospora caninum. Clin Microbiol Rev 20, 323-367.
Duret, L., 2001. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet 17, 172-175.
Fedorova, L., Fedorov, A., 2003. Introns in gene evolution. Genetica 118, 123-131.
Gabriel, H.B., de Azevedo, M.F., Palmisano, G., Wunderlich, G., Kimura, E.A., Katzin, A.M., Alves, J.M., 2015. Single-target high-throughput transcription analyses reveal high levels of alternative splicing present in the FPPS/GGPPS from Plasmodium falciparum. Sci Rep 5, 18429.
Gardner, M.J., Bishop, R., Shah, T., de Villiers, E.P., Carlton, J.M., Hall, N., Ren, Q., Paulsen, I.T., Pain, A., Berriman, M., Wilson, R.J., Sato, S., Ralph, S.A., Mann, D.J., Xiong, Z., Shallom, S.J., Weidman, J., Jiang, L., Lynn, J., Weaver, B., Shoaibi, A., Domingo, A.R., Wasawo, D., Crabtree, J., Wortman, J.R., Haas, B., Angiuoli, S.V., Creasy, T.H., Lu, C., Suh, B., Silva, J.C., Utterback, T.R., Feldblyum, T.V., Pertea, M., Allen, J., Nierman, W.C., Taracha, E.L., Salzberg, S.L., White, O.R., Fitzhugh, H.A., Morzaria, S., Venter, J.C., Fraser, C.M., Nene, V., 2005. Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. Science 309, 134-137.
Gardner, M.J., Williamson, D.H., Wilson, R.J., 1991. A circular DNA in malaria parasites encodes an RNA polymerase like that of prokaryotes and chloroplasts. Mol Biochem Parasitol 44, 115-123.
Gilbert, W., 1985. Genes-in-pieces revisited. Science 228, 823-824.
Gondim, L.F., Laski, P., Gao, L., McAllister, M.M., 2004. Variation of the internal transcribed spacer 1 sequence within individual strains and among different strains of Neospora caninum. J Parasitol 90, 119-122.
Gonzalez-Warleta, M., Castro-Hermida, J.A., Calvo, C., Perez, V., Gutierrez-Exposito, D., Regidor-Cerrillo, J., Ortega-Mora, L.M., Mezo, M., 2018. Endogenous transplacental transmission of Neospora caninum during successive pregnancies across three generations of naturally infected sheep. Vet Res 49, 106.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2013. A review of the infection, genetics, and evolution of Neospora caninum: from the past to the present. Infect Genet Evol 13, 133-150.
Grigg, M.E., Bonnefoy, S., Hehl, A.B., Suzuki, Y., Boothroyd, J.C., 2001. Success and virulence in Toxoplasma as the result of sexual recombination between two distinct ancestries. Science 294, 161-165.
Guo, Z.G., Johnson, A.M., 1995. Genetic comparison of Neospora caninum with Toxoplasma and Sarcocystis by random amplified polymorphic DNA-polymerase chain reaction. Parasitol Res 81, 365-370.
Guy, C.S., Ellis, J., Williams, D.J.L., Smith, A.J., Trees, A.J., 2005. Vaccination against

Neospora-associated abortion in cattle, Proceedings of the 20th International Conference of the World Association for the Advancement of Veteriniary Parasitology, Christchurch.
Holmdahl, J., Bjorkman, C., Stenlund, S., Uggla, A., Dubey, J.P., 1997. Bovine Neospora and Neospora caninum: One and the same. Parasitol Today 13, 40-41.
Howe, C.J., 1992. Plastid origin of an extrachromosomal DNA molecule from Plasmodium, the causative agent of malaria. J Theor Biol 158, 199-205.
Howe, D.K., Sibley, L.D., 1995. Toxoplasma gondii comprises three clonal lineages: correlation of parasite genotype with human disease. J Infect Dis 172, 1561-1566.
Huynh, M.H., Boulanger, M.J., Carruthers, V.B., 2014. A conserved apicomplexan microneme protein contributes to Toxoplasma gondii invasion and virulence. Infection and immunity 82, 4358-4368.
Innes, E.A., Wright, S.E., Maley, S., Rae, A., Schock, A., Kirvar, E., Bartley, P., Hamilton, C., Carey, I.M., Buxton, D., 2001. Protection against vertical transmission in bovine neosporosis. International journal for parasitology 31, 1523-1534.
Iriko, H., Jin, L., Kaneko, O., Takeo, S., Han, E.T., Tachibana, M., Otsuki, H., Torii, M., Tsuboi, T., 2009. A small-scale systematic analysis of alternative splicing in Plasmodium falciparum. Parasitol Int 58, 196-199.
Irimia, M., Roy, S.W., 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Res 36, 1703-1712.
Jewett, T.J., Sibley, L.D., 2003. Aldolase forms a bridge between cell surface adhesins and the actin cytoskeleton in apicomplexan parasites. Mol Cell 11, 885-894.
Jimenez-Pelayo, L., Garcia-Sanchez, M., Vazquez, P., Regidor-Cerrillo, J., Horcajo, P., Collantes-Fernandez, E., Blanco-Murcia, J., Gutierrez-Exposito, D., Roman-Trufero, A., Osoro, K., Benavides, J., Ortega-Mora, L.M., 2019. Early Neospora caninum infection dynamics in cattle after inoculation at mid-gestation with high (Nc-Spain7)- or low (Nc-Spain1H)-virulence isolates. Vet Res 50, 72.
Jung, C., Lee, C.Y., Grigg, M.E., 2004. The SRS superfamily of Toxoplasma surface proteins. International journal for parasitology 34, 285-296.
Kern, S., Agarwal, S., Huber, K., Gehring, A.P., Strodke, B., Wirth, C.C., Brugl, T., Abodo, L.O., Dandekar, T., Doerig, C., Fischer, R., Tobin, A.B., Alam, M.M., Bracher, F., Pradel, G., 2014. Inhibition of the SR protein-phosphorylating CLK kinases of Plasmodium falciparum impairs blood stage replication and malaria transmission. PLoS One 9, e105732.
Khan, A., Dubey, J.P., Su, C., Ajioka, J.W., Rosenthal, B.M., Sibley, L.D., 2011a. Genetic analyses of atypical Toxoplasma gondii strains reveal a fourth clonal lineage in North America. International journal for parasitology 41, 645-655.
Khan, A., Miller, N., Roos, D.S., Dubey, J.P., Ajzenberg, D., Darde, M.L., Ajioka, J.W., Rosenthal, B., Sibley, L.D., 2011b. A monomorphic haplotype of chromosome Ia is associated with widespread success in clonal and nonclonal populations of Toxoplasma gondii. MBio 2, e00228-00211.

Khan, A., Taylor, S., Ajioka, J.W., Rosenthal, B.M., Sibley, L.D., 2009. Selection at a single locus leads to widespread expansion of Toxoplasma gondii lineages that are virulent in mice. PLoS Genet 5, e1000404.

Khan, A., Taylor, S., Su, C., Mackey, A.J., Boyle, J., Cole, R., Glover, D., Tang, K., Paulsen, I.T., Berriman, M., Boothroyd, J.C., Pfefferkorn, E.R., Dubey, J.P., Ajioka, J.W., Roos, D.S., Wootton, J.C., Sibley, L.D., 2005. Composite genome map and recombination parameters derived from three archetypal lineages of Toxoplasma gondii. Nucleic Acids Res 33, 2980-2992.

Kim, K., Weiss, L.M., 2004. Toxoplasma gondii: the model apicomplexan. International journal for parasitology 34, 423-432.

Liddell, S., Jenkins, M.C., Collica, C.M., Dubey, J.P., 1999. Prevention of vertical transfer of Neospora caninum in BALB/c mice by vaccination. The Journal of parasitology 85, 1072-1075.

Lindsay, D.S., Dubey, J.P., 1989. In vitro development of Neospora caninum (Protozoa: Apicomplexa) from dogs. The Journal of parasitology 75, 163-165.

Lorenzi, H., Khan, A., Behnke, M.S., Namasivayam, S., Swapna, L.S., Hadjithomas, M., Karamycheva, S., Pinney, D., Brunk, B.P., Ajioka, J.W., Ajzenberg, D., Boothroyd, J.C., Boyle, J.P., Darde, M.L., Diaz-Miranda, M.A., Dubey, J.P., Fritz, H.M., Gennari, S.M., Gregory, B.D., Kim, K., Saeij, J.P., Su, C., White, M.W., Zhu, X.Q., Howe, D.K., Rosenthal, B.M., Grigg, M.E., Parkinson, J., Liu, L., Kissinger, J.C., Roos, D.S., Sibley, L.D., 2016. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. Nat Commun 7, 10147.

Lovett, J.L., Howe, D.K., Sibley, L.D., 2000. Molecular characterization of a thrombospondin-related anonymous protein homologue in Neospora caninum. Mol Biochem Parasitol 107, 33-43.

Lynch, M., Richardson, A.O., 2002. The evolution of spliceosomal introns. Curr Opin Genet Dev 12, 701-710.

Maley, S.W., Buxton, D., Rae, A.G., Wright, S.E., Schock, A., Bartley, P.M., Esteban-Redondo, I., Swales, C., Hamilton, C.M., Sales, J., Innes, E.A., 2003. The pathogenesis of neosporosis in pregnant cattle: inoculation at mid-gestation. J Comp Pathol 129, 186-195.

Manger, I.D., Hehl, A.B., Boothroyd, J.C., 1998. The surface of Toxoplasma tachyzoites is dominated by a family of glycosylphosphatidylinositol-anchored antigens related to SAG1. Infect Immun 66, 2237-2244.

Marsh, A.E., Barr, B.C., Sverlow, K., Ho, M., Dubey, J.P., Conrad, P.A., 1995. Sequence analysis and comparison of ribosomal DNA from bovine Neospora to similar coccidial parasites. J Parasitol 81, 530-535.

Marsh, A.E., Howe, D.K., Wang, G., Barr, B.C., Cannon, N., Conrad, P.A., 1999. Differentiation of Neospora hughesi from Neospora caninum based on their immunodominant surface antigen, SAG1 and SRS2. International journal for parasitology 29, 1575-1582.

McInnes, L.M., Irwin, P., Palmer, D.G., Ryan, U.M., 2006. In vitro isolation and

characterisation of the first canine Neospora caninum isolate in Australia. Veterinary parasitology 137, 355-363.

Mercier, C., Adjogble, K.D., Daubener, W., Delauw, M.F., 2005. Dense granules: are they key organelles to help understand the parasitophorous vacuole of all apicomplexa parasites? International journal for parasitology 35, 829-849.

Mercier, C., Dubremetz, J.F., Rauscher, B., Lecordier, L., Sibley, L.D., Cesbron-Delauw, M.F., 2002. Biogenesis of nanotubular network in Toxoplasma parasitophorous vacuole induced by parasite proteins. Mol Biol Cell 13, 2397-2409.

Michelin, A., Bittame, A., Bordat, Y., Travier, L., Mercier, C., Dubremetz, J.F., Lebrun, M., 2009. GRA12, a Toxoplasma dense granule protein associated with the intravacuolar membranous nanotubular network. International journal for parasitology 39, 299-306.

Miller, C., Quinn, H., Ryce, C., Reichel, M.P., Ellis, J.T., 2005. Reduction in transplacental transmission of Neospora caninum in outbred mice by vaccination. International journal for parasitology 35, 821-828.

Miller, C.M., Quinn, H.E., Windsor, P.A., Ellis, J.T., 2002. Characterisation of the first Australian isolate of Neospora caninum from cattle. Australian veterinary journal 80, 620-625.

Oladele, T.O., Sadiku, J.S., Bewaji, C.O., 2011. *In silico* characterisation of some hypothetical proteins in the proteome of *Plasmodium falciparum*. Centrepoint J. 17, 129-139.

Pedraza-Diaz, S., Marugan-Hernandez, V., Collantes-Fernandez, E., Regidor-Cerrillo, J., Rojo-Montejo, S., Gomez-Bautista, M., Ortega-Mora, L.M., 2009. Microsatellite markers for the molecular characterization of Neospora caninum: application to clinical samples. Veterinary parasitology 166, 38-46.

Preston, M.D., Campino, S., Assefa, S.A., Echeverry, D.F., Ocholla, H., Amambua-Ngwa, A., Stewart, L.B., Conway, D.J., Borrmann, S., Michon, P., Zongo, I., Ouedraogo, J.B., Djimde, A.A., Doumbo, O.K., Nosten, F., Pain, A., Bousema, T., Drakeley, C.J., Fairhurst, R.M., Sutherland, C.J., Roper, C., Clark, T.G., 2014. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun 5, 4052.

Regidor-Cerrillo, J., Diez-Fuertes, F., Garcia-Culebras, A., Moore, D.P., Gonzalez-Warleta, M., Cuevas, C., Schares, G., Katzer, F., Pedraza-Diaz, S., Mezo, M., Ortega-Mora, L.M., 2013. Genetic diversity and geographic population structure of bovine Neospora caninum determined by microsatellite genotyping analysis. PLoS One 8, e72678.

Regidor-Cerrillo, J., Pedraza-Diaz, S., Gomez-Bautista, M., Ortega-Mora, L.M., 2006. Multilocus microsatellite analysis reveals extensive genetic diversity in Neospora caninum. J Parasitol 92, 517-524.

Reichel, M.P., Alejandra Ayanegui-Alcerreca, M., Gondim, L.F., Ellis, J.T., 2013. What is the global economic impact of Neospora caninum in cattle - the billion dollar question. International journal for parasitology 43, 133-142.

Reichel, M.P., Ellis, J., 2006. If control of Neospora caninum infection is technically feasible does it make economic sense? Veterinary parasitology 142, 23-34.
Reichel, M.P., Ellis, J.T., 2002. Control options for Neospora caninum infections in cattle--current state of knowledge. N Z Vet J 50, 86-92.
Reichel, M.P., Ellis, J.T., Dubey, J.P., 2007. Neosporosis and hammondiosis in dogs. J Small Anim Pract 48, 308-312.
Reid, A.J., Vermont, S.J., Cotton, J.A., Harris, D., Hill-Cawthorne, G.A., Konen-Waisman, S., Latham, S.M., Mourier, T., Norton, R., Quail, M.A., Sanders, M., Shanmugam, D., Sohal, A., Wasmuth, J.D., Brunk, B., Grigg, M.E., Howard, J.C., Parkinson, J., Roos, D.S., Trees, A.J., Berriman, M., Pain, A., Wastling, J.M., 2012. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy. PLoS Pathog 8, e1002567.
Reynolds, S.M., Kall, L., Riffle, M.E., Bilmes, J.A., Noble, W.S., 2008. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol 4, e1000213.
Ribeiro, C.M., Soares, I.R., Mendes, R.G., de Santis Bastos, P.A., Katagiri, S., Zavilenski, R.B., de Abreu, H.F.P., Afreixo, V., 2019. Meta-analysis of the prevalence and risk factors associated with bovine neosporosis. Trop Anim Health Prod 51, 1783-1800.
Robbe, D., Passarelli, A., Gloria, A., Di Cesare, A., Capelli, G., Iorio, R., Traversa, D., 2016. Neospora caninum seropositivity and reproductive risk factors in dogs. Exp Parasitol 164, 31-35.
Rojo-Montejo, S., Collantes-Fernandez, E., Blanco-Murcia, J., Rodriguez-Bertos, A., Risco-Castillo, V., Ortega-Mora, L.M., 2009a. Experimental infection with a low virulence isolate of Neospora caninum at 70 days gestation in cattle did not result in foetopathy. Vet Res 40, 49.
Rojo-Montejo, S., Collantes-Fernandez, E., Regidor-Cerrillo, J., Alvarez-Garcia, G., Marugan-Hernandez, V., Pedraza-Diaz, S., Blanco-Murcia, J., Prenafeta, A., Ortega-Mora, L.M., 2009b. Isolation and characterization of a bovine isolate of Neospora caninum with low virulence. Veterinary parasitology 159, 7-16.
Rome, M.E., Beck, J.R., Turetzky, J.M., Webster, P., Bradley, P.J., 2008. Intervacuolar transport and unique topology of GRA14, a novel dense granule protein in Toxoplasma gondii. Infect Immun 76, 4865-4875.
Romero, J.J., Perez, E., Frankena, K., 2004. Effect of a killed whole Neospora caninum tachyzoite vaccine on the crude abortion rate of Costa Rican dairy cows under field conditions. Veterinary parasitology 123, 149-159.
Saeij, J.P., Boyle, J.P., Coller, S., Taylor, S., Sibley, L.D., Brooke-Powell, E.T., Ajioka, J.W., Boothroyd, J.C., 2006. Polymorphic secreted kinases are key virulence factors in toxoplasmosis. Science 314, 1780-1783.
Schock, A., Innes, E.A., Yamane, I., Latham, S.M., Wastling, J.M., 2001. Genetic and biological diversity among isolates of Neospora caninum. Parasitology 123, 13-23.

Sibley, L.D., Boothroyd, J.C., 1992. Virulent strains of Toxoplasma gondii comprise a single clonal lineage. Nature 359, 82-85.

Slapeta, J.R., Modry, D., Kyselova, I., Horejs, R., Lukes, J., Koudela, B., 2002. Dog shedding oocysts of Neospora caninum: PCR diagnosis and molecular phylogenetic approach. Veterinary parasitology 109, 157-167.

Spencer, J.A., Witherow, A.K., Blagburn, B.L., 2000. A random amplified polymorphic DNA polymerase chain reaction technique that differentiates between Neospora species. J Parasitol 86, 1366-1368.

Sreekumar, C., Hill, D.E., Fournet, V.M., Rosenthal, B.M., Lindsay, D.S., Dubey, J.P., 2003. Detection of Hammondia heydorni-like organisms and their differentiation from Neospora caninum using random-amplified polymorphic DNA-polymerase chain reaction. J Parasitol 89, 1082-1085.

Stenlund, S., Bjorkman, C., Holmdahl, O.J., Kindahl, H., Uggla, A., 1997. Characterization of a Swedish bovine isolate of Neospora caninum. Parasitol Res 83, 214-219.

Syn, G., Blackwell, J.M., Jamieson, S.E., Francis, R.W., 2018. An in silico pipeline to filter the Toxoplasma gondii proteome for proteins that could traffic to the host cell nucleus and influence host cell epigenetic regulation. Mem Inst Oswaldo Cruz 113, e170471.

Talevich, E., Kannan, N., 2013. Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. BMC evolutionary biology 13, 117.

Trees, A.J., Williams, D.J., 2005. Endogenous and exogenous transplacental infection in Neospora caninum and Toxoplasma gondii. Trends Parasitol 21, 558-561.

Walsh, C.P., Vemulapalli, R., Sriranganathan, N., Zajac, A.M., Jenkins, M.C., Lindsay, D.S., 2001. Molecular comparison of the dense granule proteins GRA6 and GRA7 of Neospora hughesi and Neospora caninum. International journal for parasitology 31, 253-258.

Walzer, K.A., Adomako-Ankomah, Y., Dam, R.A., Herrmann, D.C., Schares, G., Dubey, J.P., Boyle, J.P., 2013. Hammondia hammondi, an avirulent relative of Toxoplasma gondii, has functional orthologs of known T. gondii virulence genes. Proc Natl Acad Sci U S A 110, 7446-7451.

Wang, J., Tang, D., Li, W., Xu, J., Liu, Q., Liu, J., 2017. A new microneme protein of Neospora caninum, NcMIC8 is involved in host cell invasion. Exp Parasitol 175, 21-27.

Wasmuth, J., Daub, J., Peregrin-Alvarez, J.M., Finney, C.A., Parkinson, J., 2009. The origins of apicomplexan sequence innovation. Genome Res 19, 1202-1213.

Wasmuth, J.D., Pszenny, V., Haile, S., Jansen, E.M., Gast, A.T., Sher, A., Boyle, J.P., Boulanger, M.J., Parkinson, J., Grigg, M.E., 2012. Integrated bioinformatic and targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of Toxoplasma virulence. MBio 3.

Weber, F.H., Jackson, J.A., Sobecki, B., Choromanski, L., Olsen, M., Meinert, T.,

Frank, R., Reichel, M.P., Ellis, J.T., 2013. On the efficacy and safety of vaccination with live tachyzoites of Neospora caninum for prevention of neospora-associated fetal loss in cattle. Clin Vaccine Immunol 20, 99-105.

Williams, D.J., Guy, C.S., Smith, R.F., Ellis, J., Bjorkman, C., Reichel, M.P., Trees, A.J., 2007. Immunization of cattle with live tachyzoites of Neospora caninum confers protection against fetal death. Infect Immun 75, 1343-1348.

Williams, D.J., Hartley, C.S., Bjorkman, C., Trees, A.J., 2009. Endogenous and exogenous transplacental transmission of Neospora caninum - how the route of transmission impacts on epidemiology and control of disease. Parasitology 136, 1895-1900.

Williamson, D.H., Gardner, M.J., Preiser, P., Moore, D.J., Rangachari, K., Wilson, R.J., 1994. The evolutionary origin of the 35 kb circular DNA of Plasmodium falciparum: new evidence supports a possible rhodophyte ancestry. Mol Gen Genet 243, 249-252.

Wilson, R.J., Denny, P.W., Preiser, P.R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D.J., Moore, P.W., Williamson, D.H., 1996. Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum. J Mol Biol 261, 155-172.

Yeoh, L.M., Goodman, C.D., Hall, N.E., van Dooren, G.G., McFadden, G.I., Ralph, S.A., 2015. A serine-arginine-rich (SR) splicing factor modulates alternative splicing of over a thousand genes in Toxoplasma gondii. Nucleic Acids Res 43, 4661-4675.

Zhang, T.T., Zhang, X., Liu, Q., Zu, J.H., Liu, J., 2019. Protective efficacy of vaccination with NcMIC3 and NcMIC8 against Neospora caninum infection in mice. Frontiers of Agricultural Science and Engineering 6, 188-196.