Contribution of introns to the species diversity associated with the apicomplexan 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

***Corresponding author**: Larissa Calarco

School of Life Sciences, University of Technology Sydney

PO Box 123, Broadway,

NSW 2007, Australia

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

P: +61 2 9514 4161

ORCID:

Larissa Calarco: 0000-0003-1911-6866

John Ellis: 0000-0001-7328-4831

Acknowledgements:

This study was completed by LC in partial fulfilment of the Ph.D. degree at UTS

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

1. Introduction

Neospora caninum is an apicomplexan parasite of veterinary and economic significance, which is the cause of substantial reproductive losses in cattle worldwide. The parasite has a heteroxenous life cycle consisting of canines as the main definitive host, and cattle the main intermediate hosts. Vertical or transplacental transmission from dam to foetus in cattle during pregnancy however, is considered the major, most important route of transmission, which strictly involves asexual reproduction (Dubey 2003; Dubey et al. 2007; Hietala and Thurmond 1999). By comparison in the closely related *Toxoplasma gondii* species, sexual reproduction only occurs in the feline definitive host (Hutchinson 1966). While initially believed to possess a strict clonal population structure (Howe and Sibley 1995; Sibley and Boothroyd 1992), further investigations identified sexual recombination as the major driving force influencing the population structure of *T. gondii*, and therefore the emergence of new virulent strains (Ajzenberg et al. 2004; Boyle et al. 2006; Grigg et al. 2001; Khan et al. 2011; Lehmann et al. 2004).

However, horizontal transmission involving sexual reproduction in canine definitive hosts is considered infrequent for *N. caninum* when compared with vertical transmission (Bjorkman et al. 1996; Davison et al. 1999b; Reichel and Ellis 2002). Initially, limited variation was reported amongst isolates of this species, through analysis of common markers such as the Nc5 repeat (Al-Qassab et al. 2010b), 18S-like ribosomal DNA (Barber et al. 1995; Holmdahl et al. 1997; Marsh et al. 1995; Stenlund et al. 1997), and the ITS-1 region (Gondim et al. 2004; Slapeta et al. 2002). In recent years, studies have reported on the extensive genetic variation present in mini- and microsatellites, and single nucleotide polymorphisms (SNPs) and indels (insertions and deletions) identified by NGS variant detection (Al-Qassab et al. 2010a; Al-Qassab et al. 2009; Atkinson et al. 1999; Basso et al. 2009; Calarco et al. 2018; Regidor-Cerrillo et al. 2013; Regidor-Cerrillo et al. 2006). This therefore begs the question of what are

the predominant contributors to diversity within *N. caninum*, and ultimately what is driving speciation? Intra- and interspecies genetic diversity is often based on the discovery of small sequence variations, such as SNPs and indels, within protein-coding genes. Furthermore, for many non-model organisms such as *N. caninum*, which can be plagued by limited robust or complete sequencing data, larger structural variants (SVs) and copy number variation (CNVs) are often ignored. This can be due to the challenges associated with identifying, annotating, and validating such variants, which span large regions and in some cases multiple genes (Scherer et al. 2007). Large SVs however, represent a significant source of genetic diversity that can have important biological consequences.

Discussion of non-coding DNA and introns has been controversial, in terms of their origin, evolution, selective advantages, and potential myriad of functions. It is suggested that rather than representing junk DNA, introns may instead contribute to evolutionary diversity through mechanisms of alternative splicing (AS), exon shuffling, recombination, mRNA surveillance, and gene expression (Duret 2001; Fedorova and Fedorov 2003; Gilbert 1985; Lynch and Richardson 2002). Furthermore, sequence polymorphisms harboured within introns are often overlooked or underappreciated, when in fact they have the potential to influence gene expression and be implicated in genotype-phenotype relationships (Cooper 2010).

Calarco *et al*. (2018) identified SNP hotspots dispersed unevenly across the *N. caninum* genome, by comparing RNA sequencing (RNA-Seq) data generated from NC-Liverpool and NC-Nowra tachyzoites, that differ significantly in pathogenicity. Variants detected within coding regions were associated with protein-protein interactions, transcription, proteolysis, and protein kinase activity, potentially implicating these polymorphic proteins and their associated processes in *N. caninum* tachyzoite virulence. In that study, a multi-locus sequence typing approach was also developed from selected polymorphic loci that elucidated an underlying genetic population structure consisting of two major clades across a total of nine isolates. This

research demonstrated the power and contribution of SNP identification in coding regions as a source of studying genetic diversity within the species. Another potential source of diversity that is yet to be investigated for this organism however, is the presence of variation in non-coding genomic regions such as introns.

The aim of this study was to extend previous variant detection studies (Calarco et al. 2018) on the genomes of two distinct *N. caninum* isolates. The new specific focus here however, is on the extent of diversity displayed within introns and the apicoplast genome, as opposed to detecting variation in protein-coding genes. This study therefore further contributes to our understanding of the intraspecies genetic diversity existing amongst isolates of this species, and the implications for important biological processes such as virulence.

2. Materials and Methods

2.1 Generating and preparing NGS data

As part of a previous study (Calarco et al. 2018), RNA-Seq data were generated for two *N. caninum* isolates, from tachyzoites grown *in vitro* in Vero host cells (NC-Liverpool and NC-Nowra), using Illumina HiSeq2000, 100 base paired-end sequencing. It should be noted that extracted RNA was DNAse treated to ensure removal of genomic DNA. NC-Liverpool infection in cattle results in foetal death, and induces severe neosporosis in mice, characterised by severe central nervous system inflammation, encephalitis, and necrosis (Atkinson et al. 1999). By comparison, NC-Nowra shows low virulence in murine models, and has hence been suggested for use as a live vaccine to prevent foetal death in cattle (Miller et al. 2002; Weber et al. 2013; Williams et al. 2007). The NC-Liverpool transcriptome assembly produced was used in this study as an in-house reference for intron identification and downstream variant discovery. The four RNA-Seq libraries containing NC-Nowra reads, consisting of both biological and technical replicates, were aligned to the transcriptome reference using TopHat version 2.1.1 (Kim et al. 2013), and sorted and indexed using SAMtools (Li et al. 2009) (SRA accession numbers SRX4526164-SRX4526167).

For this study, whole genome sequencing (WGS) data was generated for NC-Liverpool and NC-Nowra tachyzoites grown *in vitro* in Vero cells. Sequencing was generated on the Illumina HiSeq2500 platform and consisted of four NC-Nowra libraries and five NC-Liverpool libraries derived from different tachyzoite passages. The raw NC-Nowra and NC-Liverpool reads were quality controlled using Trim Galore (version 0.5.0; www.bioinformatics.babraham.ac.uk/projects/trim_galore/), to retain reads longer than 35 bp, and bases with quality scores >25. The NC-Nowra reads were subsequently aligned to the *N. caninum* reference genome available from ToxoDB (Gajria et al. 2008) Release 39 using Bowtie2 (Langmead and Salzberg 2012) (version 2.2.8) (SRA: SRX5650793-SRX650796).

Alternatively, the NC-Liverpool reads were aligned to the Vero genome (assembly accession GCA_000409795.2) using Bowtie2, to produce unmapped FASTQ files void of host cell reads, for downstream genome assembly (SRA: SRX5650584-SRX5650588).

2.2 Identifying and annotating coding transcripts within the *de novo* transcriptome

Multiple tools were used for the annotation of the NC-Liverpool transcriptome to predict coding transcripts, identify protein domains, assign gene ontology (GO), and perform BLAST analysis (Altschul et al. 1990; Camacho et al. 2009).

TransDecoder is a tool in the Trinity software package (Haas et al. 2013) that identifies coding regions within assembled contigs or transcripts. It subsequently reports open reading frames (ORFs) that encode sequences with properties consistent with coding transcripts. Using the NC-Liverpool transcriptome, the 'TransDecoder.Predict' script was run to obtain both a FASTA file containing the coding transcript sequences, and a FASTA file containing the protein sequences corresponding to the predicted coding regions within the transcripts. The protein sequences for each transcript were then blasted against the *N. caninum* annotated proteins dataset available on ToxoDB (version 39; www.toxodb.org), using NCBI's command line BLASTP tool (Altschul et al. 1990; Camacho et al. 2009).

The NC-Liverpool transcriptome was also submitted to the FunctionAnnotator webserver (Chen et al. 2017), which is a web-based tool that offers efficient annotation of transcriptomes through sequence homology, GO, and protein domain identification. Lastly, TRAPID (Van Bel et al. 2013) was used to provide further information on the NC-Liverpool transcriptome, including the detection of ORFs, and whether or not they contained a start and/or stop codon, as well as the presence of InterProScan protein domains, and the grouping of transcripts into reference gene families. The NC-Liverpool transcriptome was searched against

the *N. caninum* reference database available through the OrthoMCL 5.0 database (Chen et al. 2006).

2.3 Identification of introns through structural variant detection in RNA-Seq data

The NC-Nowra BAM files containing RNA-Seq reads as aligned to the NC-Liverpool *de novo* transcriptome (section 2.1), were used to detect large variants and evidence for AS between the transcriptomes of these two isolates. This was addressed by implementing both the Pindel (Ye et al. 2009) and BreakDancer (Fan et al. 2014) structural variant detection algorithms. Justification for using the de novo transcriptome as opposed to the *N. caninum* reference genome sequence for this analysis, was based on the lack of robust and complete gene features and annotations for this non-model organism (Goodswen et al. 2015). For the Pindel workflow, SAMtools was first used to calculate the insert length of reads in the aligned BAM files for each NC-Nowra sample. This information was used to create a configuration file, subsequently utilised by Pindel to identify structural variation. The Pindel script was run twice to specify different minimum thresholds of supporting reads (M=10 and M=40), as opposed to the default setting of one minimum read to report a variant. For BreakDancer, the 'bam2cfg.pl' perl script was initially run to generate the necessary read group statistics in a configuration file. The 'breakdancer-max' script was subsequently run to predict SVs, based on read pairs that are mapped with discordant distances or read pair orientation. Default parameters were used, including a minimum alternative mapping quality of q=35, and an output score filter of y=40. A description of the Pindel and BreakDancer workflows is contained within Online Resource 1.

Variation in transcripts between the two isolates, as detected by both algorithms from RNA-Seq data, was visualised using the Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al. 2013), to evaluate the confidence and quality of the variants identified by both algorithms.

The visual assessment of hundreds of random variant calls assisted in the prioritisation and selection of high confidence predicted variants for laboratory validation, described below.

2.4 Laboratory identification of introns identified *in-silico*

PCR primers were designed to capture a set of four large introns, which were identified by both Pindel and BreakDancer (section 2.3; Online Resource 1, Table 1). Through BLAST analysis against annotated genome and transcriptome reference data (genome assembly accession GCA_000208865) (Reid et al. 2012), three of these were shown to represent introns within respective gene sequences, corresponding to the intron size, and start and end position within the gene. The other was allocated to a region just after the predicted open reading frame of the corresponding gene sequence in ToxoDB. The PCRs were prepared using reagents in a MyTaq Bioline PCR kit, as described by Calarco *et al.* (2018). Each reaction was accompanied by a negative control where DNA template was substituted with ddH_2O. PCRs were performed using genomic DNA extracted from NC-Liverpool and NC-Nowra cultures using the solvent extraction technique described in Calarco *et al*. (2018). This was followed by electrophoresis on 2.5% agarose gels containing GelRed, which was run at 65 V for approximately four hours.

PCR products for both NC-Liverpool and NC-Nowra were subsequently excised from gels and purified using a Qiagen QIAquick Gel Extraction kit, in accordance with the manufacturer's instructions. Sequencing was conducted on an ABI capillary sequencer by Macrogen (South Korea), and performed twice in both the forward and reverse direction. The returned ABI files were analysed using SeqTrace (Stucky 2012), and CAP3 (Huang and Madan 1999) was used to assemble the forward and reverse sequences into contigs. The final contigs from each PCR amplicon for NC-Liverpool and NC-Nowra were aligned with Clustal Omega (Sievers and Higgins 2014) for comparison.

205 **2.5 Identifying SNPs within non-coding genomic regions, using WGS data**

206 For the WGS data, the NC-Nowra reads were aligned to the NC-Liverpool ToxoDB reference
207 genome, and used to identify SNPs using the VarScan software (Koboldt et al. 2013). In
208 summary, the groomed, aligned NC-Nowra reads for each sample (BAM files; section 2.1)
209 were first sorted and indexed using SAMtools. These files were passed through the SAMtools
210 'mpileup' function, the output of which was then piped to VarScan for SNP calling. Lastly, the
211 SNP callset was filtered using VarScan's 'bam-readcount' and 'fpfilter.pl' scripts.

212 Following variant detection with VarScan, the VariantAnnotation (Obenchain et al.
213 2014) and GenomicFeatures (Lawrence et al. 2013) Bioconductor packages available in R
214 (www.R-project.org) were used to elucidate the location of SNPs within the *N. caninum*
215 genome. The published *N. caninum* GFF file from ToxoDB (version 41) containing gene
216 annotations was first used to make a TxDb object. The 'locateVariants' function was then used
217 to allocate the SNPs passing VarScan's filters within a VCF file, to either coding, intron,
218 intergenic, 3′ untranslated, 5′ untranslated, promoter, or splice site regions. The variants located
219 within introns were then extracted and randomly viewed in IGV to assess their quality.

220

221 **2.6 SNP distribution and density, and prioritisation of polymorphic regions**

222 Using VCFtools, the SNP density (SNPs/kilobase (kb)) within 50 kb genomic windows was
223 calculated for the filtered SNP callset produced by VarScan. Furthermore, the density of SNPs
224 in introns was specifically calculated and ranked (SNPs/intron length), by extracting intron
225 regions from the published *N. caninum* GFF file and cross-referencing this with the position of
226 SNPs in introns, assigned by the VariantAnnotation R package. Genes within prioritised 50 kb
227 windows, specifically those ranked with a high intron SNP density, were then annotated using
228 ToxoDB and InterProScan (Jones et al. 2014).

The SNPs previously placed within coding regions by the 'locateVariants' function were also investigated, to reveal whether genes containing a high density of SNPs within their introns, also presented a high density of SNPs in their coding regions. Additionally, the 'predictCoding' function was used to determine amino acid changes, and whether or not these SNPs represented synonymous or non-synonymous mutations. To aid these analyses, a Circos plot (Krzywinski et al. 2009) was generated to visualise and compare the location of SNPs allocated to intron and coding regions along the *N. caninum* genome, and additionally the distribution of nonsynonymous and synonymous SNPs within those coding regions.

2.7 Assembly and variant analysis of the apicoplast genome

The FASTQ files containing unmapped, paired-end NC-Liverpool WGS reads produced by Bowtie2 (i.e. host cell reads removed; section 2.1), were used for genome assembly with both ABySS (Simpson et al. 2009) and SPAdes (Bankevich et al. 2012). To optimise the assemblies, ABySS and SPAdes were run multiple times using different k-mer values (k=32, 64, 96, and 128, and k=21, 33, 55, 77, and 99 respectively). To subsequently extract the apicoplast sequence from the assembled scaffolds produced by either assembler, NCBI's command line BLASTN software (Altschul et al. 1990; Camacho et al. 2009) was used to BLAST the genome assemblies against *T. gondii* apicoplast reference resources (GenBank accession numbers U87145.2 and KE138841). The single contig assembled by ABySS representing the NC-Liverpool apicoplast, based on its alignment to the *T. gondii* reference and it being of similar length, was extracted and used as a reference for variant analysis. The prepared groomed NC-Nowra genome reads were first aligned to the apicoplast reference FASTA file, and SNP calling was subsequently performed by VarScan as previously outlined. See Online Resource 1 for further details.

253 The NC-Liverpool apicoplast sequence was annotated by blasting apicoplast reference
254 sequences of related Coccidia including *T. gondii, Hammondia hammondi, Sarcocystis*
255 *neurona,* and *Cystoisospora suis*. These apicoplast sequences were obtained from ToxoDB,
256 and aligned using command line BLASTN. Artemis (Carver et al. 2012) and DNAPlotter
257 (Carver et al. 2009) were subsequently used to graphically present an annotated apicoplast
258 genome for *N. caninum*.

3. Results

3.1 Annotation of the NC-Liverpool transcriptome

A total of 15,066 NC-Liverpool assembled transcripts predicted to be protein-coding ORFs by TransDecoder, were assigned a gene accession number following BLASTP analysis against the published *N. caninum* annotated proteins dataset. The difference between the number of transcripts assembled by Trinity (Calarco et al. 2018), and those that were predicted to be protein-coding by TransDecoder, may represent lowly expressed transcripts that prove difficult to assemble, or be a consequence of multiple isoforms being produced at single loci due to AS events (Conesa et al. 2016; Martin and Wang 2011; Smith-Unna et al. 2016). Erroneous or incomplete contigs assembled by such *de novo* tools can also result from sequencing base-calling errors, sequencing coverage, and variability within and between isolates or cultures. Alternatively, novel transcripts may be represented in the *de novo* transcriptome assembly that are not part of the annotated protein datasets for *N. caninum*. Approximately a quarter of the total transcripts were found to contain protein domains, and almost a third mapped to GO terms. From combining multiple annotation tools, Online Resource 2 summaries the main features of the NC-Liverpool transcriptome used in this study, pertinent to ORFs, BLAST hits, GOs, and protein domains, which were taken into consideration when prioritising and selecting predicted SVs for confirmation and functional analysis.

3.2 Detection and visualisation of introns in isolates

There were a total of 4,480 large variants classified as deletions, predicted by Pindel from NC-Nowra RNA-Seq reads of varying sizes (1-1452 bp), with at least ten supporting reads aligned to the NC-Liverpool reference. Increasing this parameter to a required 40 reads to call a variant however, resulted in a final callset of 1,539 predicted *in-silico* deletions. This final callset included 315 deletions greater than 50 bp, and 1,186 deletions that were one or two base pairs

284 long. The BreakDancer pipeline identified 503 large variants classified as deletions, the
285 smallest deletion being 184 bp and the largest being 1383 bp. When visualising large deletions
286 identified by both algorithms in IGV, read coverage plots showed distinct decreases in the
287 number of reads spanning the variant regions, even dropping to almost no reads covering some
288 loci. Viewing random predicted variants with IGV confirmed that increasing the number of
289 supporting reads required for Pindel to report a variant (default M=1), was conducive to
290 detecting accurate variants of high confidence.

291 To annotate the locations of large deletions predicted within their respective genes,
292 various *in-silico* deletions were blasted against annotated *N. caninum* transcripts using the
293 BLAST tool integrated in ToxoDB. Through BLASTN analysis against annotated genome and
294 transcriptome reference data (genome assembly accession GCA_000208865) (Reid et al.
295 2012), it was discovered that many predicted deletion lengths, and their start and end positions,
296 coincided with intron sequences present in the corresponding gene, confirming the predicted
297 gene annotations published in ToxoDB (Online Resource 3). In other cases, the deletions were
298 located before or after a gene (i.e. in non-coding or intergenic regions), where the length of the
299 transcriptome contig sequence for NC-Liverpool exceeded that of the predicted gene or mRNA
300 sequence provided in ToxoDB.

301 Of the introns investigated, the corresponding gene lengths and structures varied, from
302 genes containing only one intron (*NCLIV_015420*) to those containing a total of 40
303 (*NCLIV_006080*), and from sequence lengths of hundreds of base pairs (*NCLIV_041090*), to
304 tens of thousands (*NCLIV_044250*). Noteworthy protein annotations for the genes containing
305 introns identified by Pindel included a putative myosin heavy chain protein (*NCLIV_003050*),
306 a protein orthologous to GRA12 in *T. gondii* (*NCLIV_007080*), a CAMP-dependent protein
307 kinase regulatory subunit protein (*NCLIV_017370*), an F-actin capping protein alpha subunit
308 (*NCLIV_003060*), and a MORN (membrane occupation and recognition nexus) repeat

containing protein (*NCLIV_006080*). Annotations for proteins where the predicted *in-silico* deletion covered a stretch of bases preceding or succeeding the start of the gene sequence, and could for example represent spliced untranslated regions (UTRs), included two genes orthologous to MIC3 (*NCLIV_010600*) and MIC8 (*NCLIV_062770*), a protein kinase (*NCLIV_050650*), a hypothetical translation initiation factor subunit protein (*NCLIV_011760*), a putative ATP synthase subunit (*NCLIV_043880*), and a hypothetical DnaJ domain-containing protein (*NCLIV_056690*).

3.3 PCR and sequencing analysis of introns

Four gene loci (*NCLIV_002550*, *NCLIV_045190*, *NCLIV_013840*, and *NCLIV_056690*), that contained introns predicted by Pindel and BreakDancer, were subject to PCR and gel electrophoresis. Gel electrophoresis and subsequent DNA sequencing analysis confirmed that all four amplified loci were identical in size for NC-Liverpool and NC-Nowra. However, the amplicons for NC-Liverpool and NC-Nowra from *NCLIV_002550* (Online Resource 1, Table 1), were much larger than predicted when subjected to gel electrophoresis, where BLAST analysis revealed the presence of three introns located at the end of the corresponding genomic sequence, which were also amplified by the designed primers.

For two loci (*NCLIV_045190* and *NCLIV_013840*), sequencing analysis corroborated the corresponding gene and intron sequence annotations in ToxoDB. For *NCLIV_056690*, the *in-silico* deletion was located after the predicted transcript sequence in ToxoDB, which may suggest that the annotation for this gene is incorrect, based on the RNA-Seq data from both isolates. Additionally, through sequencing analysis, three SNPs in the NC-Nowra sequence were identified when compared to NC-Liverpool for this locus.

3.4 Correlation of SNP-dense introns and coding regions in WGS data

A total of 1,712 high-confidence, filtered SNPs were identified by VarScan, using WGS data from NC-Nowra and the NC-Liverpool reference genome. The majority of SNPs initially detected failed VarScan's filtering parameters based on strand bias and low supporting variant allele read counts. When visualised in IGV, detected SNPs were present across hundreds of supporting reads from all four input libraries, asserting their accuracy. The majority of SNPs were found to be located on chromosome XI (FR823392), followed by chromosome VI (FR823387) and chromosome X (FR823391). A text file based on VarScan's native output, listing the 1,712 SNPs passing the false positive filtering parameters, is contained within Online Resource 4.

The 'locateVariants' function from the VariantAnnotation package identified 501 SNPs within introns, 371 SNPs within coding regions, 840 SNPs within intergenic regions, and one SNP within a splice site (*NCLIV_041590*). Figure 1 (A-D) presents an assortment of SNPs identified by VarScan from various loci that were found to be located within intron sequences. As part of these datasets, there were a total of 19 SNPs in introns and nine SNPs in coding regions, located within large genomic contigs that are not assembled into one of the 14 *N. caninum* chromosomes. The 'predictCoding' function subsequently identified 223 SNPs within a coding region as non-synonymous, and the remaining 148 coding SNPs as synonymous.

A Circos plot is presented in Figure 2, comparing the locations of intron and coding SNPs (categorised as either nonsynonymous and synonymous mutations) in the context of the *N. caninum* reference genome. While there is an even, low distribution of SNPs across most chromosomes, there is a distinct clustering of polymorphisms on chromosomes VI (FR823387) and XI (FR823392), for SNPs located in both introns and CDS regions. Furthermore, this figure clearly demonstrates a strong correlation between and overlap of genes identified as containing SNP-dense introns and coding sequences.

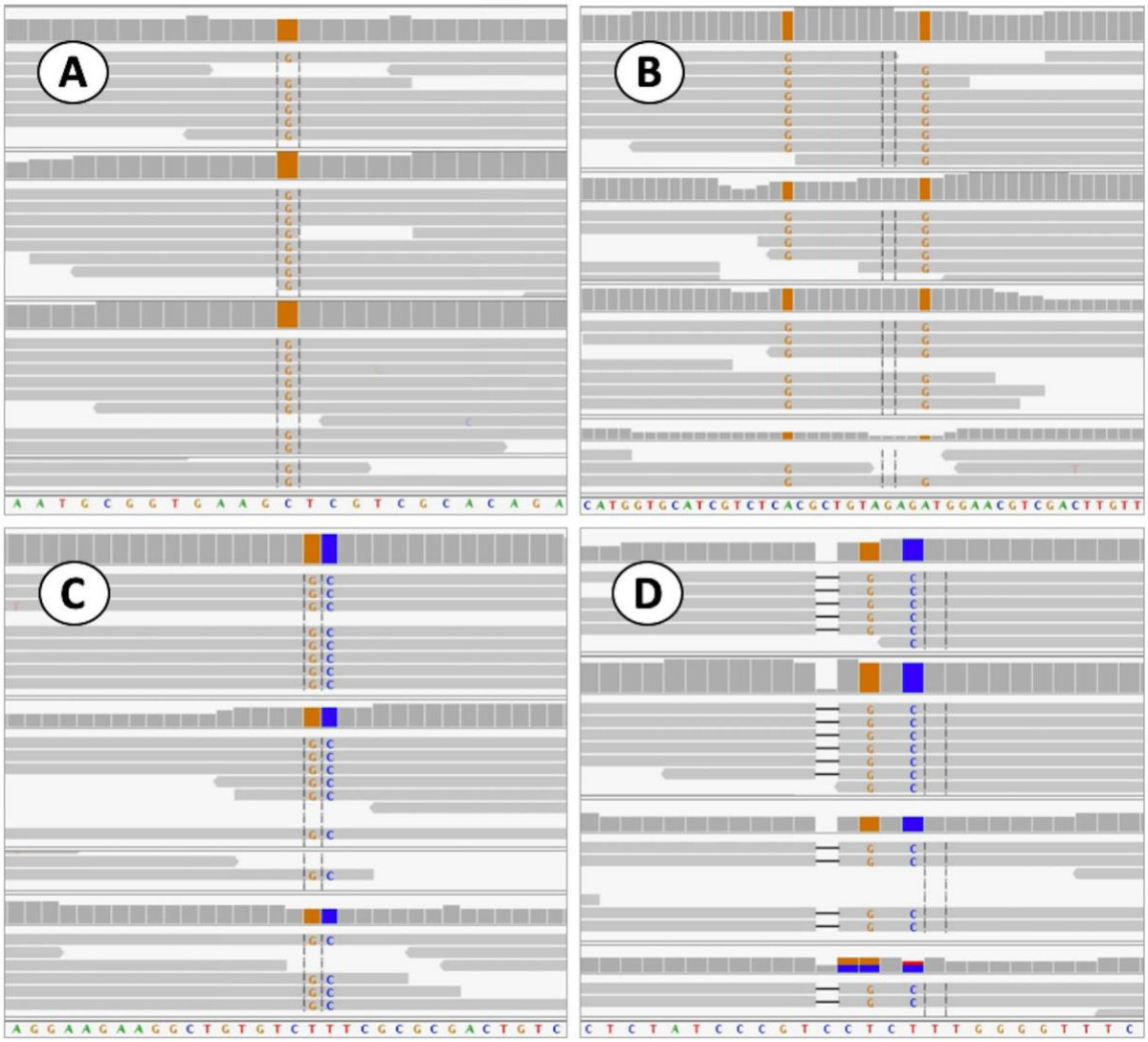


Fig. 1 Various SNPs identified by VarScan and visualised in IGV, that were subsequently allocated to introns within *N. caninum* genes. Many of the SNPs detected *in-silico* and subsequently visualised were present in all four input read libraries and across many reads, a testament to their quality and confidence. (A) Chromosome FR823381, position 622837, (B) FR823387, positions 335447 & 335458, (C) FR823389, positions 1421819 & 1421820, and (D) FR823392, positions 3761292 & 3761294

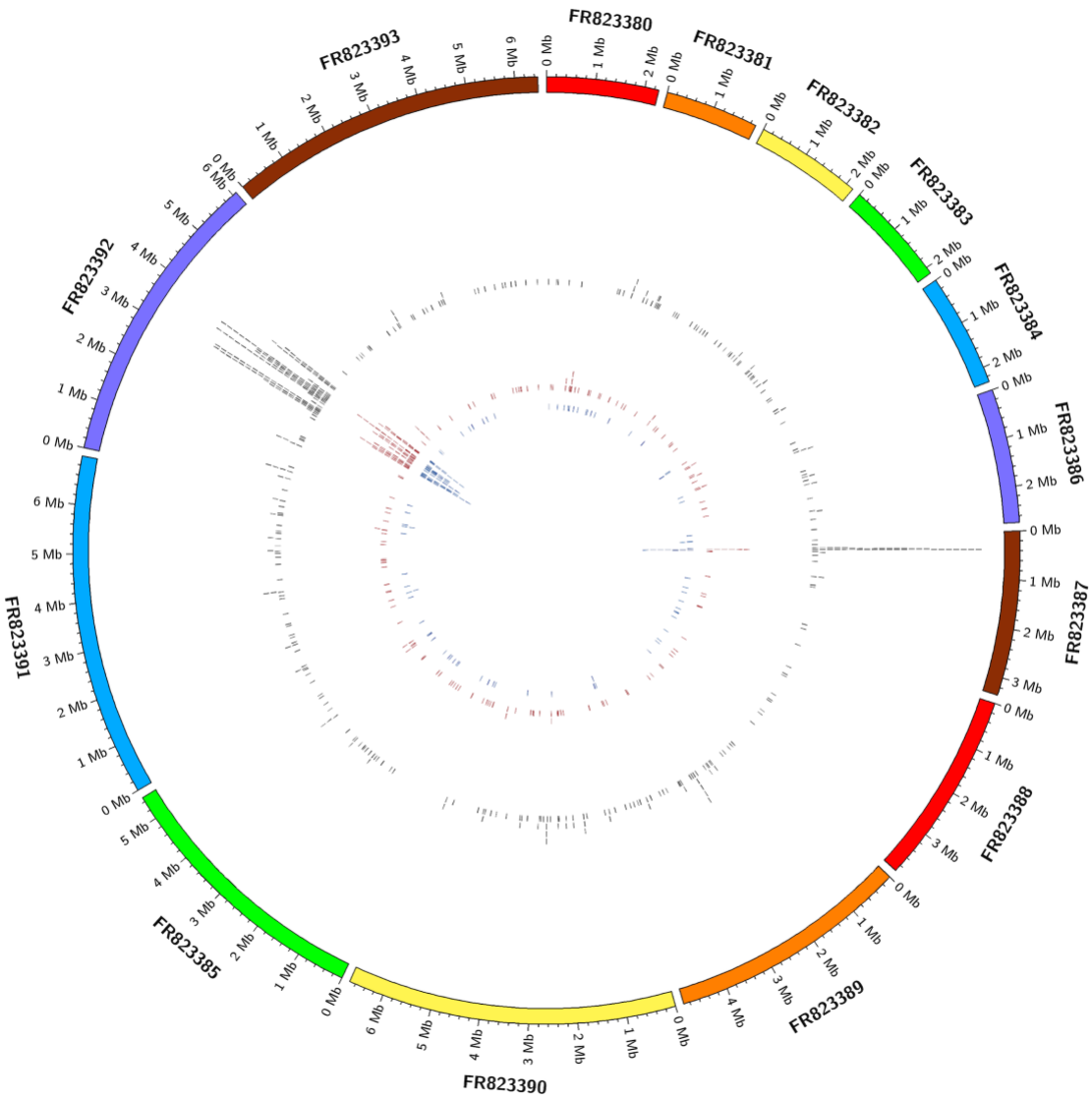


Fig. 2 A Circos plot presenting the SNPs identified by VarScan, in the context of the *N. caninum* genome. The outer track is an ideogram representing the 14 *N. caninum* chromosomes and their sizes. The middle track shows the distribution of non-coding SNPs located within introns along the genome, where each tile represents one SNP. The innermost track is split into two plots, collectively representing SNPs identified within coding gene regions: the outermost track (red) depicts non-synonymous SNPs, whereas the inner track (blue) depicts synonymous SNPs

3.5 Prioritisation of introns and genomic regions, based on SNP density

While the average SNP density across chromosome VI was 0.04 SNPs/50 kb, there were two 50 kb genomic windows (300-400 kb) with considerably higher SNP densities, which contained a total of 101 SNPs (including 56 SNPs in introns and 25 in coding regions). Within this genomic window, there were eight *N. caninum* genes and three tRNA encoding genes. These annotated genes included a GDP-mannose 4, 6 dehydratase (*NCLIV_015780*), a putative fatty acyl-CoA desaturase involved in calcium ion binding (*NCLIV_015970*), and a mitochondrial carrier protein (*NCLIV_015850*). For chromosome XI, while the average SNP density was 0.06 SNPs/50kb, there were 15 50kb genomic windows containing >15 SNPs and a SNP density $\geq$0.32 SNPs/50 kb. In total there were 126 genes within these regions, with a collective 519 SNPs (including 184 SNPs in introns and 150 coding SNPs). Annotated genes of interest included 26S proteasome regulatory subunit protein (*NCLIV_056700*), a protein kinase-like superfamily protein (*NCLIV_056620*), an ABC transporter (*NCLIV_057400*), a 50S ribosome binding GTPase protein (*NCLIV_057390*), a ribosomal biogenesis family protein (*NCLIV_057180*), and a WD domain, G-beta repeat containing protein (*NCLIV_057900*).

With respect to SNP density within individual introns investigated along chromosomes VI and XI, most introns contained either one or two SNPs, and a maximum of five SNPs were identified in only two introns. On chromosome VI, the gene with the highest SNP density was *NCLIV_015820*, which is a calcium-activated potassium channel protein with a signal peptide and transmembrane domain. While this is a large protein, there are eight introns present, four of which contained SNPs called by VarScan. This was closely followed by *NCLIV_015830*, an ABC transporter type 1 protein with transmembrane domains and a P-loop containing nucleoside triphosphate hydrolase domain. This protein is involved in transmembrane transport, ATP binding, and ATPase activity, where 14 of its 32 introns contained SNPs. For chromosome XI, *NCLIV_057570* had the highest SNP density for one of its introns, which is

401 an RNA polymerase II-associated protein involved in transcription. For this gene, three of four
402 of its introns were found to contain SNPs, the sizes of which were quite small compared to the
403 surrounding exons. The next highest SNP-dense intron belonged to *NCLIV_058710*, encoding
404 a superoxide dismutase and associated with oxidation-reduction and superoxide metabolic
405 processes, and metal ion binding. This gene is only 857 bp in length, with two short introns.
406 When ranked, some of the genes appeared multiple times at the top of the SNP density
407 list, as they harboured multiple introns, each of which contained SNPs called by VarScan. For
408 chromosome VI, these gene products included the ABC transporter involved in transmembrane
409 transport, ATP binding, and ATPase activity (*NCLIV_015830*), a glutamic-acid rich protein
410 (*NCLIV_015800*), and the large transmembrane, ion channel protein (*NCLIV_015820*). For
411 chromosome XI, this included an intron-binding protein aquarius, which is a pre-mRNA
412 splicing factor (*NCLIV_057470*), a hypothetical protein belonging to the kinesin-like protein
413 family involved in microtubule movement (*NCLIV_056770*), and another hypothetical protein
414 with a BSD domain found in transcription factors and synapse-associated proteins
415 (*NCLIV_057450*). Many proteins within the prioritised SNP density windows, harbouring
416 SNPs in both coding and intron sequences, also had BLASTP hits to serine-rich adhesin for
417 platelets, GPI-anchored proteins, and gel-forming secreted mucins. Online Resource 5 contains
418 a tabulated list of genes of interest within prioritised polymorphic hotspots on chromosomes
419 VI and XI, containing introns ranked with a high SNP density. The annotations provided,
420 including domains, superfamilies, and GOs, were collated from InterProScan and ToxoDB
421 records.
422 A summary of the methodology and main results for the comparative transcriptome and
423 genome analyses between NC-Liverpool and NC-Nowra, is presented in Figure 3.

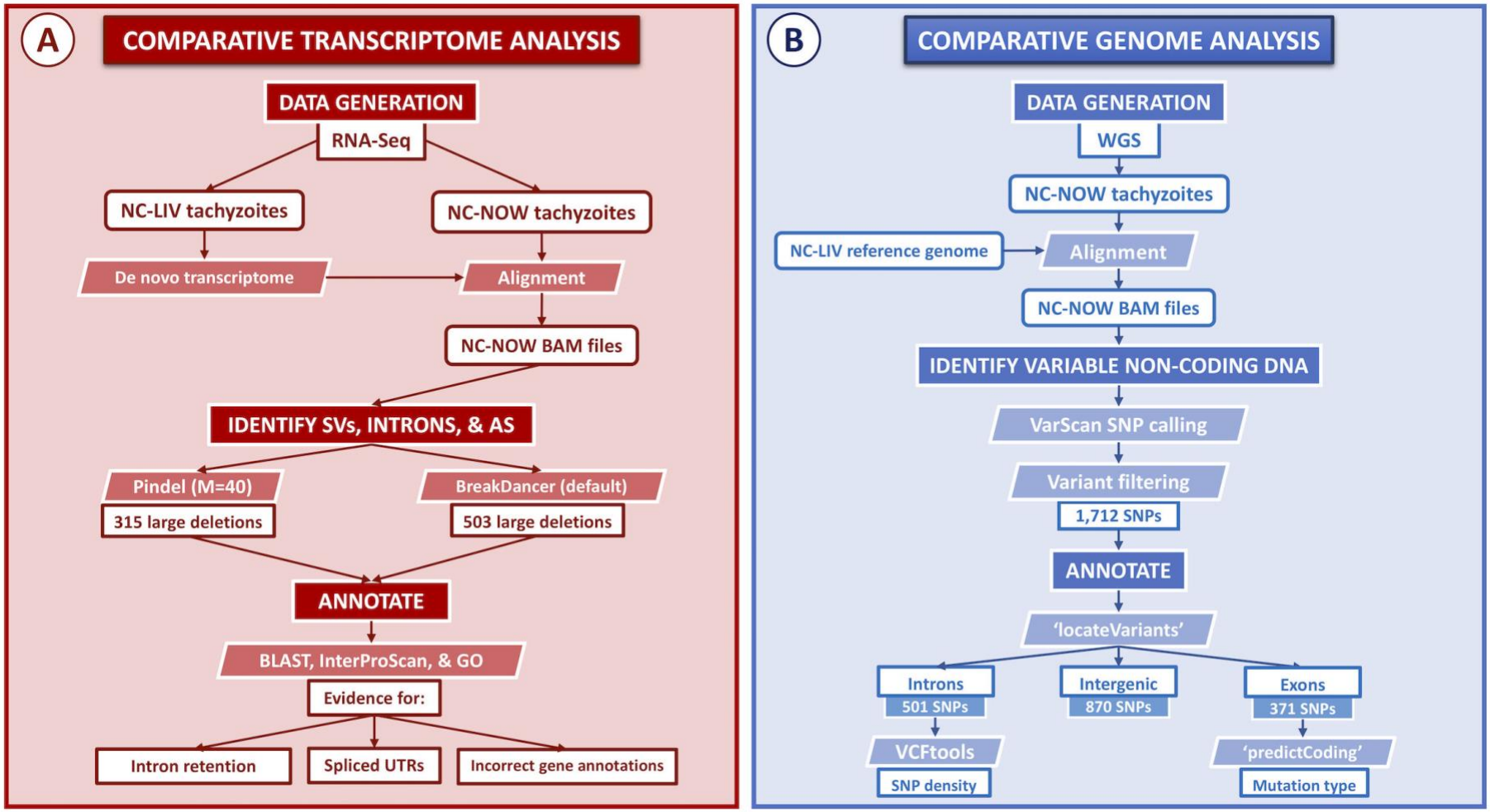


Fig. 3 Summary of the methodology and main findings pertaining to the comparative transcriptome and genome analyses. (A) Groomed RNA-Seq reads generated from NC-Nowra (NC-NOW) tachyzoites were aligned to a de novo NC-Liverpool (NC-LIV) reference transcriptome,

to aid in the discovery of structural variants (SVs), and mechanisms of alternative splicing (AS) such as intron retention. For this, the Pindel and BreakDancer algorithms were employed to identify variable transcripts between the two isolates, where the results were subsequently annotated using various tools and databases, to assign biological context. (B) Whole genome sequencing (WGS) was performed for the NC-Nowra isolate, where groomed reads were aligned to the current published NC-Liverpool reference genome. Subsequently, the VarScan software was implemented to identify variable non-coding DNA sequences along the *N. caninum* genome, where identified single nucleotide polymorphisms (SNPs) were filtered to obtain a high confidence callset. Lastly, the SNPs were annotated to elucidate their genomic location, distribution, and function, followed by the prioritisation of SNP-dense intron-containing genes

3.6 Apicoplast sequence assembly and variation within non-nuclear DNA

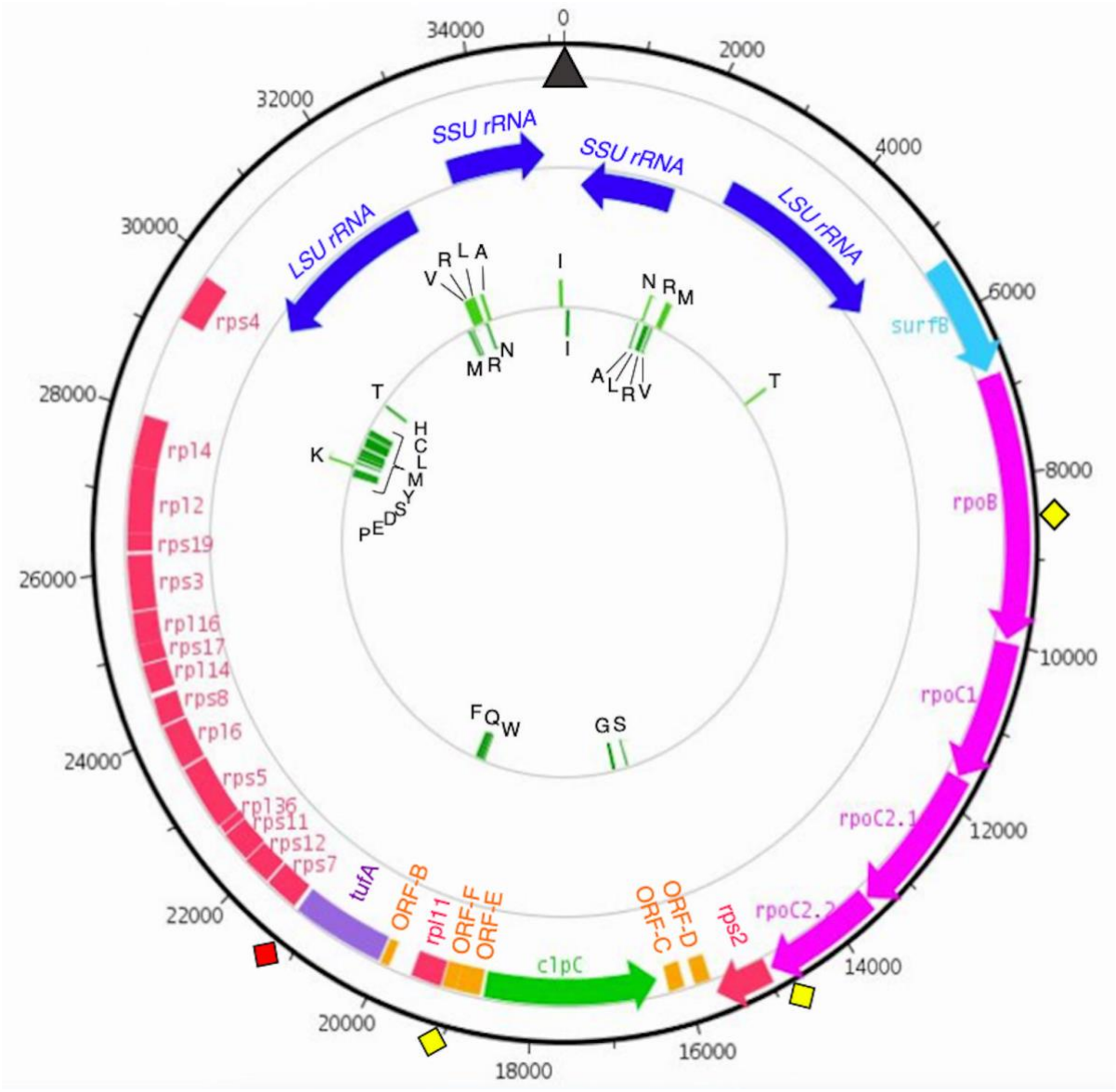
Using a k-mer length of 96, ABySS assembled a contig of length 35,163 bp, that had high confidence BLAST hits to two *T. gondii* apicoplast reference sequences. The genome assemblies produced by SPAdes and the other k-mer lengths specified for ABySS, failed to produce a single contig similar in length to the *T. gondii* reference apicoplast genome, and to what has been previously estimated for *N. caninum* (Gleeson and Johnson 1999).

The BLAST results to related Coccidia included hits to *SSU rRNA*, *LSU rRNA*, RNA polymerase β subunit (*rpoB*), and RNA polymerase β' subunits (*rpoC1* and *rpoC2*), as well as numerous tRNAs and ribosomal proteins. The BLASTN results of apicoplast sequences from various *T. gondii* isolates investigated, generated percentage identities >93%, and e-values of 0.0 for many genes. The alignments to other Coccidia including *H. hammondi*, *C. suis*, and *S. neurona*, also demonstrated high sequence similarity for the length of the apicoplast sequence. There were a total of 29 protein coding genes annotated, including those encoding five putative open reading frames (ORFs) and 17 ribosomal proteins, as well as 33 tRNAs, and two large and two small subunit rRNA coding genes. The *N. caninum* apicoplast is also AT-rich at 78.4%, which is similar to what has been reported for other related species (Arisue and Hashimoto 2015; Wilson et al. 1996).

As described for other Coccidia and *Plasmodium* species, *N. caninum* also appears to contain an inverted repeat region that includes duplicated small and large subunit rRNAs and nine tRNAs, positioned head-to-head. Additionally, approximately half of the apicoplast genes are transcribed in a clockwise direction, and the other half in a counter-clockwise direction. For $tRNA^{Glu}$ (position 28,068-28,140), $tRNA^{Leu}$ (position 28,730-28,997) and ribosomal protein S8 (*rps8*; position 24,099-24,453), the *N. caninum* sequences only aligned to *S. neurona*, where these sequences were not part of the reference genomes used to conduct BLAST analysis for the other Coccidia investigated. There were also 33 in-frame 'TGA' stop

459 codons detected for a total of 19 protein-coding genes, as well as ‘TAG’ and ‘TAA’ in-frame
460 stop codons in *rpoB* and *rpl16* respectively. With respect to stop codon usage for the 29 protein
461 coding genes, a total of 26 CDSs ended in ‘TAA’, compared to only three genes containing a
462 ‘TAG’ stop codon (ORF-D, *rpl4*, and *rps5*). Figure 4 presents the annotated apicoplast genome
463 for *N. caninum* based on BLAST analysis, the results of which are available in Online Resource
464 6.

465 Following variant analysis of the NC-Liverpool apicoplast sequence against the NC-
466 Nowra genome reads, there were a total of three SNPs and one insertion across the ~35 kb
467 sequence. These SNPs were located within *rpoB* (position 8,574), *rpoC2.2* (position 14,866),
468 and within ORF-F (position 19,102). The one indel identified by VarScan was an insertion
469 spanning three bases at position 21,243, just before the start codon of the *tufA* gene. While the
470 insertion does not disrupt the initiation ‘ATG’ codon site, whether or not the presence and
471 location of this variant affects the translation of this gene, is unclear. The VarScan variants
472 described are also marked on Figure 4.



473

474 **Fig. 4 Schematic of the annotated *N. caninum* apicoplast sequence assembled with ABySS,**

475 **based on genome data from NC-Liverpool.**

476 The start of the apicoplast sequence indicated by the grey arrowhead, which also represents
477 position '0' on the figure, marks the joining point of the two inverted repeat regions. This
478 includes duplicated small and large subunit rRNAs (*SSU rRNA* and *LSU rRNA* respectively)
479 displayed on the middle track, and nine tRNAs arranged head to head presented in green on
480 the innermost track. The outermost track contains the CDS regions of the annotated apicoplast
481 genes, where those transcribed clockwise are shown on the outer layer (denoted by the
482 arrowheads), and those transcribed counter-clockwise are shown on the inner layer of this

track. The RNA polymerase genes are presented in magenta, and are all transcribed clockwise, and the putative open reading frames (ORF-D, ORF-C, ORF-E, and ORF-F) are coloured orange and are all transcribed counter-clockwise. Ribosomal proteins are coloured light red on the outermost track, and tRNAs are labelled and presented in green on the inner-most circle. The three SNPs detected by VarScan between the *N. caninum* NC-Liverpool and NC-Nowra isolates are labelled with a yellow diamond, and the three base pair indel is labelled by the red diamond

4. Discussion

Neospora caninum is a protozoan parasite that presents a major economic and veterinary threat to beef and dairy industries on a global scale. However, although the parasite was first identified over 30 years ago, currently there are no treatment or prevention options against neosporosis that are both efficacious and affordable. Sexual recombination is considered rare in the parasite's life cycle, which is instead dominated by transplacental transmission in intermediate hosts, involving asexual reproduction (Dubey 2003; Dubey et al. 2007; Hietala and Thurmond 1999). This begs the question, of what are the driving forces contributing to diversity within the species, both at the genetic and phenotypic levels, and how can elucidating such diversity contribute to the development of vaccines and drugs against the disease?

The presence and impact of large SVs has been studied in various Protozoa, where for example the clustering of SVs around hypervariable gene families and within virulence factor genes in *Plasmodium falciparum* (Claessens et al. 2014) and *T. gondii* (Boyle et al. 2008; Khan et al. 2009) respectively, are associated with antigenic diversity and adaptive evolution in these species. In this study, structural variant calling pipelines identified hundreds of large deletions exceeding 50 bp, using RNA-Seq data derived from NC-Liverpool and NC-Nowra tachyzoites. However, upon investigation of these large deletions identified *in-silico*, most represented introns. Many of the introns were found to be located within hypothetical proteins, however many contained transmembrane domains and/or signal peptides, and included GOs for protein binding, oxidation-reduction process, proteolysis, metal ion binding, RNA processing, methyltransferase activity, F-actin capping protein complex, aminopeptidase activity, and cAMP-dependent protein kinase complex. The investigated transcripts and their corresponding genes differed in both the number of introns they harboured, and the lengths of their respective protein, gene, and intron sequences (Online Resource 3). These results suggest that introns are not being correctly removed by post-transcriptional processes in isolates of this species, where

instead introns are being retained in mRNA sequences. However, this idea requires further investigation to elucidate whether such mechanisms of AS are operating in *N. caninum* which seems likely based on this evidence.

There were also large deletions predicted *in-silico* that appeared to be located before or after a gene (i.e. in non-coding or intergenic regions), some of which contained important annotations for known apicomplexan virulence factors such as microneme proteins (MICs) and protein kinases (Carruthers and Sibley 1997; Cerede et al. 2002; Cerede et al. 2005; Kessler et al. 2008; Khan et al. 2009; Lim et al. 2012; Saeij et al. 2006). Obtained from transcriptome data, these results suggest that the *in-silico* deletions may be located within UTRs or regulatory regions of these genes, and may therefore influence expression, or alternatively that reference resources for *N. caninum* available on ToxoDB are incomplete or inaccurate.

Furthermore in this study, variant analysis using WGS data from these two isolates revealed a distinct clustering of SNPs within specific genomic regions on chromosomes VI (FR823387) and XI (FR823392). This was in stark contrast to the rest of the *N. caninum* genome, which appeared mostly monomorphic, presenting limited polymorphisms scattered along the remaining chromosomes (Figure 2). These results reflect the distribution of SNPs reported by Calarco *et al*. (2018) using RNA-Seq data from the NC-Liverpool and NC-Nowra isolates, within tachyzoite associated, transcriptionally active protein-coding genes.The SNP callset identified by VarScan in this study consisted of 1,712 high-confidence SNPs, which were categorised by their location (i.e. those located within introns and exons), as well as whether coding SNPs represented non-synonymous and synonymous mutations. Interestingly, the number of non-synonymous SNPs (223), was greater than synonymous SNPs (148) within coding regions, which may suggest that these identified polymorphic “hotspot” regions are or have been subjected to different selective evolutionary pressures. There was also a high number

539 of SNPs identified within intergenic regions (840), almost equating to the number of SNPs
540 located within exons (371) and introns (501) combined.

541 By combining the datasets obtained, a set of genes ranked based on SNP density across
542 their introns (SNPs/intron length) on chromosomes VI and XI were investigated and annotated
543 (Online Resource 5). Within this list of prioritised SNP-dense introns and their corresponding
544 genes, there were three genes from both chromosomes VI and XI that contained multiple
545 introns ranked highest according to density. Classification of SNP-dense intron-containing
546 genes by GO terms, revealed recurring terms such as ATP binding, ATPase activity,
547 transmembrane transport, and various transcription and translation processes. Collectively, the
548 identified hypervariable genes and their annotated functions could represent contributors to *N.*
549 *caninum* diversity, and therefore important parasite phenotypes such as virulence between
550 isolates.

551 Since its initial discovery, many studies have investigated and reported on the
552 biological and genetic diversity characterising isolates of *N. caninum* using a range of
553 techniques and molecular targets (Al-Qassab et al. 2010a; Al-Qassab et al. 2009; Atkinson et
554 al. 1999; Calarco et al. 2018; Davison et al. 1999a; Marsh et al. 1995; Pedraza-Diaz et al. 2009;
555 Regidor-Cerrillo et al. 2013; Regidor-Cerrillo et al. 2006; Schock et al. 2001). While diversity
556 in coding sequences is a powerful source for studying genetic diversity, additional research is
557 investigating the importance of variation in non-coding genomic regions such as introns.
558 Rather than being disregarded as non-coding, "junk" DNA, the positions, sequences, and
559 composition of introns have the potential to address a range of genomic and evolutionary
560 questions (Irimia and Roy 2008). Alternative splicing events have also been identified and
561 studied in unicellular eukaryotes with comparatively small genomes in protozoan parasites,
562 where organisms including *P. falciparum* and *T. gondii* are estimated to have 50-75% of their
563 genes interrupted by introns (Gardner et al. 2002; Suvorova and White 2014). Many of the

documented alternatively spliced transcripts of protozoan parasites code for proteins that are also categorised as potential vaccine or drug candidates (Agarwal et al. 2011; Gabriel et al. 2015; Kern et al. 2014). Collectively, evidence suggests that AS is prevalent in apicomplexan parasites, and has the potential to produce diverse, biologically significant transcripts that differ throughout life cycle stages, where it is estimated that approximately 22% and 16% of protein-coding genes undergo AS in *T. gondii* (Yeoh et al. 2015) and *Plasmodium* species (Iriko et al. 2009) respectively.

The number of non-coding (intron and intergenic) SNPs identified throughout the *N. caninum* genome in this study, in comparison to SNPs located within exons, appears to reflect the fundamental principles contained in Kimura's neutral theory of evolution (Kimura 1991). However, the presence of non-coding intron sequences in eukaryotic genomes remains enigmatic, as they must be correctly removed through splicing following transcription, if the function of the gene product is to be preserved, which is also metabolically expensive (Berget et al. 1977; Chow et al. 1977). It is understood that introns are therefore generally more amenable and tolerable to sequence mutations in many species, thus representing a source for estimating neutral mutation rates within and between populations (Gilbert 1978; Irimia and Roy 2008). Furthermore, the necessary elimination of introns from transcripts does not strictly suggest that they are exempt from selection (Bergman and Kreitman 2001; Jareborg et al. 1999; Llopart et al. 2002), where they can become part of coding regions in alternatively spliced transcripts, be involved in transcription and translation regulation (Beaulieu et al. 2011; Bianchi et al. 2009; Chapman and Walter 1997), and the preservation of pre-mRNA secondary structure (Kirby et al. 1995).

Lastly, variant analysis conducted on the apicoplast sequence assembled in this study for NC-Liverpool, demonstrated that this organelle is highly conserved between the two isolates investigated. Only three SNPs were identified and one insertion present across the ~35

589 kb sequence, between bases 8,557-21,245. It is understood that the content of apicoplast
590 genomes across species is mostly conserved, encoding small and large subunit ribosomal
591 RNAs (*SSU rRNA* and *LSU rRNA*), subunits of bacterial RNA polymerases, an elongation
592 factor, and a member of the clp family of molecular chaperones (*clpC*), plus a complete set of
593 tRNAs, and 16-18 ribosomal proteins (Sato 2011; Wilson et al. 1996; Wilson and Williamson
594 1997). The apicoplast mainly encodes housekeeping genes (Wilson et al. 1996), although
595 inhibition or disruption of the apicoplast results in parasite death, attributing this organelle as
596 vital to parasite survival and viability (Dahl and Rosenthal 2008; Fichera and Roos 1997;
597 Goodman and McFadden 2013). Consequently, the apicoplast has become an attractive target
598 for therapeutic drug design.

599 Based on the almost identical apicoplast sequences investigated for NC-Liverpool and
600 NC-Nowra in this study, and the demonstrated crucial function of the apicoplast, it is highly
601 likely that this organelle is under strong selective pressure within and between species. This
602 was also demonstrated by the high sequence similarity between the *N. caninum* apicoplast
603 assembly generated, and reference sequences from other closely related Coccidia including *T.*
604 *gondii* (≥93%), *H. hammondi* (≥89%), *S. neurona* (≥75%), and *C. suis* (≥79%) (Online
605 Resource 6). The gene catalogue, function, and arrangement, as well as base composition and
606 nucleotide sequence, also appear to be extremely similar amongst and conserved between these
607 related Coccidia, as well as *Plasmodium* species (Figure 4). The results of this study suggest
608 that the *N. caninum* apicoplast also contains an inverted repeat region described for other
609 Apicomplexa, consisting of the small and large subunit rRNAs and nine tRNAs, as well as
610 genes transcribed in both a clockwise and counter-clockwise direction. Furthermore, there were
611 a total of 33 in-frame 'TGA' stop codons identified across 19 apicoplast ORFs for *N. caninum*.
612 This is consistent with what has been reported for *T. gondii* apicoplast gene sequences, where
613 these codons are predicted to encode tryptophan (Wilson 2002).

By investigating SNP density within nuclear DNA, and comparing these results for non-coding and coding SNPs, this study alludes to the possibility that forces of natural selection are acting on and influencing these biologically important regions. This is especially true of chromosomes VI and XI, where both this study and the results presented by Calarco *et al.* (2018), suggest that the polymorphic hotspots within these small genomic regions alone have and are being subjected to evolutionary selective pressures. This may indicate that the transcriptionally active genes within these hotspots are instrumental in, or at least associated with, the reported pathogenic variability between isolates of this species. As only two isolates were investigated in this study, such research should be extended to sequence prioritised genes of interest and their identified polymorphisms for additional isolates. This would elucidate if and what types of selective pressures may be contributing to the intraspecies genetic and phenotypic diversity reported for *N. caninum*, within both coding and non-coding genomic regions. Future efforts should also be dedicated to further validating and extending the investigation of AS, and specifically intron retention, between multiple isolates of this species. However, the limited amount of robust, accurate, and annotated data for non-model organisms such as *N. caninum* and related Coccidia including *Neospora hughesi* and *Hammondia heydorni*, continues to plague and hinder the execution of such analyses to contribute to the existing body of knowledge. Furthermore, the effect of SNPs within introns should be investigated, as well as the presence and retention of introns themselves within genes, including the potential influence of introns on gene expression and regulation.

5. Conclusions

This study took advantage of RNA-Seq data generated from tachyzoites of two isolates of *N. caninum*, that differ in their pathogenicity in murine models (Calarco et al. 2018). Structural variant detection pipelines identified hundreds of large deletions present in NC-Nowra reads, as aligned to an NC-Liverpool reference transcriptome. Analysis of these deletions revealed that many represented introns or spliced UTRs within respective gene sequences. Furthermore, variant analysis was performed using whole genome data from the two isolates, which confirmed the presence of polymorphic hotspots on chromosomes VI (FR823387) and XI (FR823392), and identified SNP-dense introns within genes of biological interest. Additionally, variant analysis conducted on the apicoplast sequence, revealed that this crucial organelle is highly conserved in structure, content, and arrangement between not only the two *N. caninum* isolates investigated, but also with other related Coccidia. This research provides additional knowledge on genetic sources contributing to the intraspecies diversity in *N. caninum*, which may have important biological impacts in areas such as virulence, population structure, and vaccine and drug targets. This research also highlights the need to not only focus on protein-coding genes when investigating population genetics, phylogenetics, and evolutionary mechanisms, but also to extend such research to non-coding genomic regions, such as introns, which have the power to expand upon our knowledge of intra- and interspecies diversity. Future efforts should be made to extend these findings to additional *N. caninum* isolates and related species, to elucidate the potential evolutionary pressures operating on and influencing polymorphic hotspots in the genome of this species.

Data Availability:

Nucleotide sequence data reported in this paper have been submitted to GenBank under accession numbers **MK754233**-**MK754238**. Whole genome sequencing data for NC-Liverpool and NC-Nowra have been deposited in SRA under accession numbers **SRX5650584**-**SRX5650588** (https://www.ncbi.nlm.nih.gov//bioproject/PRJNA531306) and **SRX5650793**-**SRX650796** (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA531328), respectively. The assembled, annotated apicoplast sequence for NC-Liverpool has been submitted to GenBank under accession number **MK770339** (https://www.ncbi.nlm.nih.gov/nuccore/MK770339).

Conflict of Interest:

The authors declare they have no competing interests

Funding:

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

References

Agarwal S et al. (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 doi:10.1002/jcb.23034

Ajzenberg D, Banuls AL, Su C, Dumetre A, Demar M, Carme B, Darde ML (2004) Genetic diversity, clonality and sexuality in Toxoplasma gondii Int J Parasitol 34:1185-1196 doi:10.1016/j.ijpara.2004.06.007

Al-Qassab S, Reichel MP, Ellis J (2010a) A second generation multiplex PCR for typing strains of Neospora caninum using six DNA targets Mol Cell Probes 24:20-26 doi:10.1016/j.mcp.2009.08.002

Al-Qassab S, Reichel MP, Ellis JT (2010b) On the biological and genetic diversity in Neospora caninum Diversity 2:1424-2818

Al-Qassab S, Reichel MP, Ivens A, Ellis JT (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

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

Arisue N, Hashimoto T (2015) Phylogeny and evolution of apicoplasts and apicomplexan parasites Parasitol Int 64:254-259 doi:10.1016/j.parint.2014.10.005

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

Bankevich A et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing J Comput Biol 19:455-477 doi:10.1089/cmb.2012.0021

Barber JS, Holmdahl OJ, Owen MR, Guy F, Uggla A, Trees AJ (1995) Characterization of the first European isolate of Neospora caninum (Dubey, Carpenter, Speer, Topper and Uggla) Parasitology 111 (Pt 5):563-568 doi:10.1017/s0031182000077039

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 160:43-50 doi:10.1016/j.vetpar.2008.10.085

Beaulieu E, Green L, Elsby L, Alourfi Z, Morand EF, Ray DW, Donn R (2011) Identification of a novel cell type-specific intronic enhancer of macrophage migration inhibitory factor (MIF) and its regulation by mithramycin Clin Exp Immunol 163:178-188 doi:10.1111/j.1365-2249.2010.04289.x

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA Proc Natl Acad Sci U S A 74:3171-3175

Bergman CM, Kreitman M (2001) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences Genome Res 11:1335-1345 doi:10.1101/gr.178701

Bianchi M, Crinelli R, Giacomini E, Carloni E, Magnani M (2009) A potent enhancer element in the 5'-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene Gene 448:88-101 doi:10.1016/j.gene.2009.08.013

Bjorkman C, Johansson O, Stenlund S, Holmdahl OJ, Uggla A (1996) Neospora species infection in a herd of dairy cattle J Am Vet Med Assoc 208:1441-1444

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 103:10514-10519 doi:10.1073/pnas.0510319103
Boyle JP, Saeij JP, Harada SY, Ajioka JW, Boothroyd JC (2008) Expression quantitative trait locus mapping of toxoplasma genes reveals multiple mechanisms for strain-specific differences in gene expression Eukaryot Cell 7:1403-1414 doi:10.1128/EC.00073-08
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 doi:10.1093/gbe/evy188
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications BMC Bioinformatics 10:421 doi:10.1186/1471-2105-10-421
Carruthers VB, Sibley LD (1997) Sequential protein secretion from three distinct organelles of Toxoplasma gondii accompanies invasion of human fibroblasts Eur J Cell Biol 73:114-123
Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data Bioinformatics 28:464-469 doi:10.1093/bioinformatics/btr703
Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization Bioinformatics 25:119-120 doi:10.1093/bioinformatics/btn578
Cerede O, Dubremetz JF, Bout D, Lebrun M (2002) The Toxoplasma gondii protein MIC3 requires pro-peptide cleavage and dimerization to function as adhesin EMBO J 21:2526-2536 doi:10.1093/emboj/21.11.2526
Cerede O, Dubremetz JF, Soete M, Deslee D, Vial H, Bout D, Lebrun M (2005) Synergistic role of micronemal proteins in Toxoplasma gondii virulence J Exp Med 201:453-463 doi:10.1084/jem.20041672
Chapman RE, Walter P (1997) Translational attenuation mediated by an mRNA intron Curr Biol 7:850-859
Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups Nucleic Acids Res 34:D363-368 doi:10.1093/nar/gkj123
Chen TW et al. (2017) FunctionAnnotator, a versatile and efficient web tool for non-model organism annotation Sci Rep 7:10430 doi:10.1038/s41598-017-10952-4
Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA Cell 12:1-8
Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, Kwiatkowski D (2014) Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis PLoS Genet 10:e1004812 doi:10.1371/journal.pgen.1004812
Conesa A et al. (2016) A survey of best practices for RNA-seq data analysis Genome Biol 17:13 doi:10.1186/s13059-016-0881-8
Cooper DN (2010) Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes Hum Genomics 4:284-288
Dahl EL, Rosenthal PJ (2008) Apicoplast translation, transcription and genome replication: targets for antimalarial antibiotics Trends Parasitol 24:279-284 doi:10.1016/j.pt.2008.03.007

Davison HC et al. (1999a) In vitro isolation of Neospora caninum from a stillborn calf in the UK Res Vet Sci 67:103-105 doi:10.1053/rvsc.1998.0272
Davison HC, Otter A, Trees AJ (1999b) Estimation of vertical and horizontal transmission parameters of Neospora caninum infections in dairy cattle Int J Parasitol 29:1683-1689
Dubey JP (2003) Review of Neospora caninum and neosporosis in animals Korean J Parasitol 41:1-16
Dubey JP, Schares G, Ortega-Mora LM (2007) Epidemiology and control of neosporosis and Neospora caninum Clin Microbiol Rev 20:323-367 doi:10.1128/CMR.00031-06
Duret L (2001) Why do genes have introns? Recombination might add a new piece to the puzzle Trends Genet 17:172-175
Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping Curr Protoc Bioinformatics 45:15 16 11-11 doi:10.1002/0471250953.bi1506s45
Fedorova L, Fedorov A (2003) Introns in gene evolution Genetica 118:123-131
Fichera ME, Roos DS (1997) A plastid organelle as a drug target in apicomplexan parasites Nature 390:407-409 doi:10.1038/37132
Gabriel HB, de Azevedo MF, Palmisano G, Wunderlich G, Kimura EA, Katzin AM, Alves JM (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 doi:10.1038/srep18429
Gajria B et al. (2008) ToxoDB: an integrated Toxoplasma gondii database resource Nucleic Acids Res 36:D553-556 doi:10.1093/nar/gkm981
Gardner MJ et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum Nature 419:498-511 doi:10.1038/nature01097
Gilbert W (1978) Why genes in pieces? Nature 271:501
Gilbert W (1985) Genes-in-pieces revisited Science 228:823-824
Gleeson MT, Johnson AM (1999) Physical characterisation of the plastid DNA in Neospora caninum Int J Parasitol 29:1563-1573
Gondim LF, Laski P, Gao L, McAllister MM (2004) Variation of the internal transcribed spacer 1 sequence within individual strains and among different strains of Neospora caninum J Parasitol 90:119-122 doi:10.1645/GE-134R
Goodman CD, McFadden GI (2013) Targeting apicoplasts in malaria parasites Expert Opin Ther Targets 17:167-177 doi:10.1517/14728222.2013.739158
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 45:305-318 doi:10.1016/j.ijpara.2015.01.006
Grigg ME, Bonnefoy S, Hehl AB, Suzuki Y, Boothroyd JC (2001) Success and virulence in Toxoplasma as the result of sexual recombination between two distinct ancestries Science 294:161-165 doi:10.1126/science.1061888
Haas BJ et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis Nat Protoc 8:1494-1512 doi:10.1038/nprot.2013.084
Hietala SK, Thurmond MC (1999) Postnatal Neospora caninum transmission and transient serologic responses in two dairies Int J Parasitol 29:1669-1676

Holmdahl J, Bjorkman C, Stenlund S, Uggla A, Dubey JP (1997) Bovine Neospora and Neospora caninum: One and the same Parasitol Today 13:40-41 doi:10.1016/s0169-4758(97)81616-x
Howe DK, Sibley LD (1995) Toxoplasma gondii comprises three clonal lineages: correlation of parasite genotype with human disease J Infect Dis 172:1561-1566
Huang X, Madan A (1999) CAP3: A DNA sequence assembly program Genome Res 9:868-877
Hutchinson WM (1966) Recent observations on the biology of Toxoplasma gondii Trans Ophthalmol Soc U K 86:185-189
Iriko H et al. (2009) A small-scale systematic analysis of alternative splicing in Plasmodium falciparum Parasitol Int 58:196-199 doi:10.1016/j.parint.2009.02.002
Irimia M, Roy SW (2008) Spliceosomal introns as tools for genomic and evolutionary analysis Nucleic Acids Res 36:1703-1712 doi:10.1093/nar/gkn012
Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs Genome Res 9:815-824
Jones P et al. (2014) InterProScan 5: genome-scale protein function classification Bioinformatics 30:1236-1240 doi:10.1093/bioinformatics/btu031
Kern S et al. (2014) Inhibition of the SR protein-phosphorylating CLK kinases of Plasmodium falciparum impairs blood stage replication and malaria transmission PLoS One 9:e105732 doi:10.1371/journal.pone.0105732
Kessler H, Herm-Gotz A, Hegge S, Rauch M, Soldati-Favre D, Frischknecht F, Meissner M (2008) Microneme protein 8--a new essential invasion factor in Toxoplasma gondii J Cell Sci 121:947-956 doi:10.1242/jcs.022350
Khan A, Dubey JP, Su C, Ajioka JW, Rosenthal BM, Sibley LD (2011) Genetic analyses of atypical Toxoplasma gondii strains reveal a fourth clonal lineage in North America Int J Parasitol 41:645-655 doi:10.1016/j.ijpara.2011.01.005
Khan A, Taylor S, Ajioka JW, Rosenthal BM, Sibley LD (2009) Selection at a single locus leads to widespread expansion of Toxoplasma gondii lineages that are virulent in mice PLoS Genet 5:e1000404 doi:10.1371/journal.pgen.1000404
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions Genome Biol 14:R36 doi:10.1186/gb-2013-14-4-r36
Kimura M (1991) The neutral theory of molecular evolution: a review of recent evidence Jpn J Genet 66:367-386
Kirby DA, Muse SV, Stephan W (1995) Maintenance of pre-mRNA secondary structure by epistatic selection Proc Natl Acad Sci U S A 92:9047-9051
Koboldt DC, Larson DE, Wilson RK (2013) Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection Curr Protoc Bioinformatics 44:15 14 11-17 doi:10.1002/0471250953.bi1504s44
Krzywinski M et al. (2009) Circos: an information aesthetic for comparative genomics Genome Res 19:1639-1645 doi:10.1101/gr.092759.109
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2 Nat Methods 9:357-359 doi:10.1038/nmeth.1923
Lawrence M et al. (2013) Software for computing and annotating genomic ranges PLoS Comput Biol 9:e1003118 doi:10.1371/journal.pcbi.1003118
Lehmann T, Graham DH, Dahl ER, Bahia-Oliveira LM, Gennari SM, Dubey JP (2004) Variation in the structure of Toxoplasma gondii and the roles of selfing, drift, and epistatic

selection in maintaining linkage disequilibria Infect Genet Evol 4:107-114 doi:10.1016/j.meegid.2004.01.007
Li H et al. (2009) The Sequence Alignment/Map format and SAMtools Bioinformatics 25:2078-2079
Lim DC, Cooke BM, Doerig C, Saeij JP (2012) Toxoplasma and Plasmodium protein kinases: roles in invasion and host cell remodelling Int J Parasitol 42:21-32 doi:10.1016/j.ijpara.2011.11.007
Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002) Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection Proc Natl Acad Sci U S A 99:8121-8126 doi:10.1073/pnas.122570299
Lynch M, Richardson AO (2002) The evolution of spliceosomal introns Curr Opin Genet Dev 12:701-710
Marsh AE, Barr BC, Sverlow K, Ho M, Dubey JP, Conrad PA (1995) Sequence analysis and comparison of ribosomal DNA from bovine Neospora to similar coccidial parasites J Parasitol 81:530-535
Martin JA, Wang Z (2011) Next-generation transcriptome assembly Nat Rev Genet 12:671-682 doi:10.1038/nrg3068
Miller CM, Quinn HE, Windsor PA, Ellis JT (2002) Characterisation of the first Australian isolate of Neospora caninum from cattle Aust Vet J 80:620-625 doi:10.1111/j.1751-0813.2002.tb10967.x
Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M (2014) VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants Bioinformatics 30:2076-2078 doi:10.1093/bioinformatics/btu168
Pedraza-Diaz S, Marugan-Hernandez V, Collantes-Fernandez E, Regidor-Cerrillo J, Rojo-Montejo S, Gomez-Bautista M, Ortega-Mora LM (2009) Microsatellite markers for the molecular characterization of Neospora caninum: application to clinical samples Vet Parasitol 166:38-46 doi:10.1016/j.vetpar.2009.07.043
Regidor-Cerrillo J et al. (2013) Genetic diversity and geographic population structure of bovine Neospora caninum determined by microsatellite genotyping analysis PLoS One 8:e72678 doi:10.1371/journal.pone.0072678
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 92:517-524 doi:10.1645/GE-713R.1
Reichel MP, Ellis JT (2002) Control options for Neospora caninum infections in cattle--current state of knowledge N Z Vet J 50:86-92 doi:10.1080/00480169.2002.36288
Reid AJ et al. (2012) Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy PLoS Pathog 8:e1002567 doi:10.1371/journal.ppat.1002567
Saeij JP et al. (2006) Polymorphic secreted kinases are key virulence factors in toxoplasmosis Science 314:1780-1783 doi:10.1126/science.1133690
Sato S (2011) The apicomplexan plastid and its evolution Cell Mol Life Sci 68:1285-1296 doi:10.1007/s00018-011-0646-1
Scherer SW et al. (2007) Challenges and standards in integrating surveys of structural variation Nat Genet 39:S7-15 doi:10.1038/ng2093
Schock A, Innes EA, Yamane I, Latham SM, Wastling JM (2001) Genetic and biological diversity among isolates of Neospora caninum Parasitology 123:13-23

Sibley LD, Boothroyd JC (1992) Virulent strains of Toxoplasma gondii comprise a single clonal lineage Nature 359:82-85 doi:10.1038/359082a0
Sievers F, Higgins DG (2014) Clustal omega Curr Protoc Bioinformatics 48:3 13 11-16 doi:10.1002/0471250953.bi0313s48
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data Genome Res 19:1117-1123 doi:10.1101/gr.089532.108
Slapeta JR, Koudela B, Votypka J, Modry D, Horejs R, Lukes J (2002) Coprodiagnosis of Hammondia heydorni in dogs by PCR based amplification of ITS 1 rRNA: differentiation from morphologically indistinguishable oocysts of Neospora caninum Vet J 163:147-154 doi:10.1053/tvjl.2001.0599
Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies Genome Res 26:1134-1144 doi:10.1101/gr.196469.115
Stenlund S, Bjorkman C, Holmdahl OJ, Kindahl H, Uggla A (1997) Characterization of a Swedish bovine isolate of Neospora caninum Parasitol Res 83:214-219 doi:10.1007/s004360050236
Stucky BJ (2012) SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms J Biomol Tech 23:90-93 doi:10.7171/jbt.12-2303-004
Suvorova ES, White MW (2014) Transcript maturation in apicomplexan parasites Curr Opin Microbiol 20:82-87 doi:10.1016/j.mib.2014.05.012
Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration Briefings in bioinformatics 14:178-192 doi:10.1093/bib/bbs017
Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes Genome Biol 14:R134 doi:10.1186/gb-2013-14-12-r134
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 20:99-105 doi:10.1128/CVI.00225-12
Williams DJ, Guy CS, Smith RF, Ellis J, Bjorkman C, Reichel MP, Trees AJ (2007) Immunization of cattle with live tachyzoites of Neospora caninum confers protection against fetal death Infect Immun 75:1343-1348 doi:10.1128/IAI.00777-06
Wilson RJ (2002) Progress with parasite plastids J Mol Biol 319:257-274 doi:10.1016/S0022-2836(02)00303-0
Wilson RJ et al. (1996) Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum J Mol Biol 261:155-172
Wilson RJ, Williamson DH (1997) Extrachromosomal DNA in the Apicomplexa Microbiol Mol Biol Rev 61:1-16
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads Bioinformatics 25:2865-2871 doi:10.1093/bioinformatics/btp394
Yeoh LM, Goodman CD, Hall NE, van Dooren GG, McFadden GI, Ralph SA (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 doi:10.1093/nar/gkv311