On the application of reverse vaccinology to parasitic diseases: some thoughts on feature selection and ranking of vaccine candidates

Stephen J. Goodswen[a], Paul J. Kennedy[b], John T. Ellis[a,*]

[a] *School of Life Sciences, University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia*

[b] *School of Software, Faculty of Engineering and Information Technology and the Centre for Artificial Intelligence, University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia*

Note: Sequence data reported in this paper are available from ??? under accession number XXX

[*]Corresponding Author. Tel.: +61 2 9514 4161.

Mailing address**:** School of Life Sciences, University of Technology Sydney (UTS), PO Box 123, Broadway NSW 2007, Australia.

Email address: John.Ellis@uts.edu.au

ABSTRACT

Reverse vaccinology has the potential to rapidly advance vaccine development against parasites, but it is unclear which features studied *in-silico* will advance vaccine development. Here we consider *Neospora caninum* which is a globally distributed protozoan parasite causing significant economic and reproductive loss to cattle industries worldwide. The aim of this study was to use a reverse vaccinology approach to compile a worthy vaccine candidate list for *N. caninum*, including pathogen-associated molecular patterns (PAMPs). The *in silico* approach essentially involved obtaining protein characteristics from public databases or computationally predicting them for every known *Neospora* protein. A wide range of data on features or attributes of the genes or proteins were collected and analysed using an automated high-throughput process. The final vaccine list compiled was judged to be the optimum within the constraints of available data, current knowledge, and existing bioinformatics programs. We consider and provide some suggestions and experience on how ranking of vaccine candidate lists can be performed. This study is therefore important in that it provides a valuable resource for establishing new directions in vaccine research against neosporosis and other parasitic diseases of economic and medical importance.

Keywords: *Neospora caninum*, *in silico* vaccine discovery, reverse vaccinology, machine learning, pathogen-associated molecular patterns.

1. Introduction

Previous studies (Goodswen et al., 2015; Goodswen et al., 2013b, 2014a) have proposed that subunit vaccine candidates against target eukaryotic pathogens can theoretically be discovered using an *in silico* approach based on the principle of reverse vaccinology. Many publications describe reverse vaccinology in detail (Davies and Flower, 2007; Donati and Rappuoli, 2013; Jones, 2012). An *in silico* approach in essence *predicts* protein antigens using biological data pertinent to the target pathogen. This is in direct contrast to the traditional culture-based approach that identifies protein antigens by cultivating and dissecting the pathogen in the laboratory. Candidacy validation in appropriate assays, animal models, and ultimately hosts is still nevertheless a mandatory requirement for both approaches. The foremost advantage of the *in silico* approach, however, is that theoretically every single pathogen protein can be computationally analysed for its vaccine candidacy potential; whereas the traditional approach is limited to those proteins captured during the laboratory process. Furthermore, predicting a list of potential antigens for laboratory validation is relatively inexpensive and takes only weeks of computing time in comparison to the more labour intensive and expensive traditional approach.

The primary biological data for the *in silico* approach are protein sequences from the target pathogen. Such sequences contain information/signals for predicting informative protein characteristics. For example, predicting whether a protein is typically found on cell membranes or secreted. As yet there is no bioinformatics program to predict that a protein will induce the desired protective immune response in a host. This is mainly due to not knowing the target characteristic to predict. In other words, no distinguishing signal within a protein sequence has so far been detected that clearly indicates a protein is immunogenic. Consequently, the primary *in silico* strategy is to predict and/or gather known protein

characteristics that suggest a protein might be immunogenic and worthy of laboratory validation.

As a case study we consider the application of reverse vaccinology to vaccine development against *Neospora caninum.* This obligate intracellular parasite of the phylum Apicomplexa is of high veterinary importance and has been reviewed in detail recently (Dubey and Schares, 2011; Goodswen et al., 2013c; Monney and Hemphill, 2014). Infection by this parasite can cause the clinical disease neosporosis resulting in abortion in cattle. The worldwide accumulative financial loss due to abortions is estimated to exceed US$1 billion annually (Reichel et al., 2013). This loss presents a substantial burden to both the dairy and beef industries. A vaccine is considered one form of control.

As to date, there are just over 10,000 *N. caninum* protein sequences available through publicly accessible databases (Goodswen et al., 2013c). It is unlikely this protein count truly represents every protein that can potentially be expressed by the *N. caninum* genome, but, in principle, the primary data to initiate reverse vaccinology exist. Determining which proteins out of the 10,000 that are worthy of laboratory validation was the end goal of this study. The foremost obstacle to a successful outcome is that it remains unclear what constitutes an ideal vaccine candidate for *N. caninum*. What is clear, and is well-supported by many studies (Andrianarivo et al., 2005; Rosbottom et al., 2008; Williams et al., 2007), is that the whole *N. caninum* organism in the form of tachyzoites induces both humoral (antibodies) and cell-mediated immunity (CMI) responses by the host in an attempt to control infection. What is not known is the type of antigen that mediates the protective immune response and whether only a limited or a large number of various antigen types are involved. Nevertheless, a possible vaccine formulation could potentially entail a cocktail of recombinant protein antigens fused with adjuvants containing appropriate pathogen-associated molecular patterns

(PAMPS). Ideally, the protein antigens will be from tachyzoites and bradyzoites as both these life cycle stages occur in cattle.

To achieve the desired immunological response and subsequent memory to *N. caninum*, dendritic cells (DCs) must present immunogenic peptides, originating from protein antigens in a vaccine, on major histocompatibility complex (MHC) class II molecules. The concept behind a PAMP-based adjuvant is to initiate DC uptake of synthetically linked protein antigens. Only a few PAMPs have been reported for apicomplexans, but this may be because protozoan PAMP identification is at an early stage in comparison to identification of bacterial and viral PAMPs (Gazzinelli and Denkers, 2006). The PAMPs identified for bacteria and viruses are not commonly found in eukaryotic organisms and are predominantly lipids and lipoproteins e.g. lipopolysaccharide (LPS), peptidoglycan, and flagellin (Kawai and Akira, 2010). The apicomplexan PAMPs most described in the literature are agonists to Toll-like receptors (TLR), for example, *T. gondii* profilin-like protein (Plattner et al., 2008). This protein was shown to activate TLR11 in mouse cells (Yarovinsky et al., 2005), but there is no known equivalent TLR11 in cattle. Another potential PAMP, observed on *T. gondii* to activate TLR4 and TLR2, is a heat-shock protein (Aosai et al., 2002; Del Rio et al., 2004). The role of these TLR agonists in phagocytosis is unclear (Kerrigan and Brown, 2009).

Cattle DCs have innate receptor molecules (Seabury et al., 2010) and it is reasonable to assume they have evolved to recognise PAMPs that are common to broad classes of pathogens including apicomplexans. Immature DCs express a large array of endocytic/phagocytic receptors (Banchereau et al., 2000) with the expectation that some will recognise and bind to conserved portions on foreign proteins. Moreover, the contribution of several receptor types is likely to be involved in this recognition. The proteins sought for a PAMP-based adjuvant will need to possess the typical properties of known molecules containing PAMPs. That is, proteins should be naturally exposed to DCs, conserved among

many apicomplexans, perform survival essential but non-virulent functions, and be dissimilar to host proteins.

The aim of this study was to use an *in silico* approach to compile a worthy vaccine candidate list for *N. caninum* including PAMPs. This involves collecting both existing and computationally predicted protein characteristics for every known *N. caninum* protein, irrespective of current annotation. Candidate choice is based on assessing each collated protein characteristic for its evidence contribution potential but primarily using antigen properties as selection criteria. The approach used is an automated high-throughput process as it is impractical to perform a case-by-case examination of each protein. The final vaccine list is judged to be the optimum candidates within the constraints of available data, current knowledge, and existing bioinformatics programs. We also consider the method of ranking of vaccine candidates and provide some suggestions on methodologies.

2. Materials and methods

The following sections describe the specific *N. caninum* protein characteristics that were collected and how these characteristics were obtained or computationally predicted. All collected characteristics are recorded in Supplementary data S1.

2.1. Neospora caninum annotation

Protein sequences for NC-Liverpool were downloaded from UniProtKB (Apweiler et al., 2004) in a FASTA format. Pertinent characteristics about each protein were also downloaded from UniProtKB and recorded. This included UniProt ID; Length (number of amino acids); Protein name; Annotation (a score from one to five that provides a heuristic measure of the annotation content of a UniProtKB entry (a five implies a well-annotated protein); Protein existence (indicates the type of evidence that supports the existence of the protein. Five types: experimental evidence at protein level, experimental evidence at transcript level, protein inferred from homology, protein predicted, and Protein uncertain); Gene Ontology (GO) term for biological process; GO term for molecular function; GO term for cellular component; Gene names; Date of last modification (date that the annotation was last modified not including the sequence); Subcellular location (description of the location of the mature protein); Post-translational modification (PTMs); and Function (describes the general function(s) of a protein). Note that in this work flow only the UniProt ID is guaranteed to have a value.

2.2. Evidence to support protein existence

Almost 99% of the NC-Liverpool protein sequences provided by UniProtKB come from the translations of predicted coding sequences (CDS). Three different gene prediction methods were utilised to provide supporting evidence that a protein exists: *Ab initio* (or

intrinsic) (Fickett, 1996), evidence based (or extrinsic) (Borodovsky et al., 1994) using ESTs, and genome sequence comparison (van Baren et al., 2002).

Two *ab initio* gene finder programs were used: AUGUSTUS (Stanke et al., 2004; Stanke et al., 2006) and GlimmerHMM (Majoros et al., 2004; Pertea and Salzberg, 2002). Both use a variation of hidden Markov models (HMMs) (Sleator, 2010) to statistically model structure of DNA sequences and each gene finder has its own complex internal algorithm to decode the HMM into gene predictions (Brent, 2007). A training dataset comprising validated genes is required for both gene finders to train HMMs. The dataset here was created with all genes from the *T. gondii* genome that have evidence for protein expression based on mass spectrometry analyses. These genes (3,432 in total) were downloaded from ToxoDB version 12. The *ab initio* gene finders predicted the genes within each of the 14 NC-Liverpool chromosomes. These chromosomes were from assembly # ASM20886v2 and downloaded from ToxoDB version 12.

The nucleotide sequences for ESTs were downloaded in a FASTA format from dbEST release 130101 (http://www.ncbi.nlm.nih.gov/dbEST/). There were 25094 ESTs for *N. caninum*. Two evidence based gene finders called BLAT (Kent, 2002) and GMAP (Wu and Watanabe, 2005) aligned the ESTs to the *N. caninum* chromosomes and predicted exon locations (including a prediction of the number of exons per gene). Both programs output the exon information in a PSL format (a format specific to BLAT). An in-house Perl script extracted the relevant data from the PSL files and concatenated nucleotide sequences from each exon member of a gene.

The genome sequence comparison method works on the principle that conserved regions between related organisms are more likely to be coding, and conversely divergent regions more likely to be non-coding. N-SCAN (Gross and Brent, 2006) was used to perform

the sequence comparison, which requires a target genome and an informant genome to help identify coding regions and splice sites. The informant genome used was the combination of *T. gondii* chromosomes downloaded from ToxoDB version 12. The output file from N-SCAN is a Gene Transfer Format (GTF).

The predicted gene sequences derived from the *ab initio* gene finders, N-SCAN, and the concatenated sequences from BLAT and GMAP were aligned to existing NC-Liverpool gene sequences using BLASTN (a program that is part of the Basic Local Alignment Search Tool (BLAST) suite of applications (Camacho et al., 2009)). The NC-Liverpool gene sequences were downloaded from ToxoDB version 12. A score was computed based on: query coverage * sequence percentage similarity – where query coverage is the percent of the predicted sequence that overlaps the existing sequence; and sequence percentage similarity is the nucleotide similarity at the same positions between the predicted and existing sequence over the length of the coverage area. The scores were reported as a value between 0 and 1, where a 1.0 represents a maximum score i.e. 100% query coverage and 100% sequence percentage similarity. Note that the concatenated sequences are in effect equivalent to mRNA sequences. The query coverage previously described was therefore computed differently to account for possible intron regions on the existing gene i.e. query coverage = (query length – number of matches) / query length.

The *ab initio* gene finders and N-SCAN also predicted exon locations. An in-house Perl script concatenated the exons to create mRNA sequences. All predicted mRNA sequences, including those from the RNA-Seq experiment (see section 2.3), were translated to amino acids. BLASTP (a program that is also part of the BLAST suite of applications (Camacho et al., 2009) was used to compare the predicted amino acid sequences with the existing NC-Liverpool proteins. A score was computed in a similar manner to that previously described for the gene comparison.

A final evidence score for each existing NC-Liverpool protein was computed by averaging the scores obtained from the three methods. In summary, these scores were obtained from BLAST comparisons from predicted AUGUSTUS genes and mRNAs; GlimmerHMM genes and mRNAs; N-SCAN genes and mRNAs; BLAT mRNAs; and GMAP mRNAs. The final score was recorded and represents a confidence level between 0 and 1 that the protein exists. For instance, a value of 1 indicates that the predicted gene and protein sequences, computed by all three methods, significantly matches the existing gene and protein sequences, and subsequently strongly supports a protein's existence.

2.3. RNA-Seq evidence

RNA-Seq data was obtained from a previous study (Goodswen et al., 2015). In brief, RNA-Seq reads were generated from three biological replicates of cultured NC-Liverpool tachyzoite populations. Tophat2 (Trapnell et al., 2012) mapped NC-Liverpool reads to the NC-Liverpool reference genome. Cufflinks (Trapnell et al., 2012) assembled the aligned RNA-Seq reads into transcripts in a General Transfer Format (GTF). An in-house Perl script was used to extract exon base pair (bp) locations and mRNA sequences from the GTF file. Furthermore, day three and four Illumina paired-end RNA-Seq reads from the original annotation study (Reid et al., 2012) were downloaded from ArrayExpress (accession number E-MTAB-549) at https://www.ebi.ac.uk/arrayexpress/. These reads were used to obtain exon locations and mRNA sequences as previously described in this section (referred to henceforth as Sanger RNA-Seq).

2.4. Identifying homologs

A stand-alone BLASTP was used to search the National Center for Biotechnology Information (NCBI) protein database called 'nr' to find the nearest homolog to every NC-

Liverpool protein. The nr database was downloaded from NCBI FTP site. This database contains all non-redundant GenBank CDS translations, NCBI RefSeq proteins, proteins from Protein Database (PDB), UniProt, International Protein Sequence Database (PIR), and Protein Research Foundation (PRF). The BLASTP 'hit' with the highest bitscore determined the nearest homolog. Seven characteristics per protein pertaining to the nearest homolog were recorded: name, weight, protein length, source organism, source gene, NCBI GI number, and percentage sequence similarity. These characteristics provided clues to the identity of uncharacterised *Neospora* proteins and/or supported or opposed current annotation. The 'weight' characteristic was a devised score from 0 to 7 to indicate the reliability of the homolog protein name as source of evidence. A protein name with the single word 'none' is designated the least reliable with a score of 7. All other scores are based on the word(s) contained within the name: unnamed protein product = 6, hypothetical protein = 5, conserved hypothetical protein = 4, hypothetical protein, conserved = 4, putative = 3, PREDICTED = 2, partial = 1 and all other protein names = 0 (most reliable).

2.5. Identifying immune-exposed proteins

Vacceed (Goodswen et al., 2014c) was used to determine the probability that a NC-Liverpool protein is naturally exposed to the immune system. The probability was computed within *Vacceed* by a set of machine learning algorithms trained on known exposed and non-exposed proteins (Goodswen et al., 2013a). *Vacceed* is essentially a configurable framework of linked programs and for this study was configured using SignalP 4.0 (Petersen et al., 2011) (predicts presence and location of signal peptide cleavage sites); WoLF PSORT 0.2 (Horton et al., 2007) and TargetP 1.1 (Emanuelsson et al., 2007) (predict subcellular localization);

TMHMM 2.0 (Krogh et al., 2001) (predicts transmembrane domains in proteins); and Phobius (Kall et al., 2004) (predicts transmembrane topology and signal peptides).

2.6. Identifying proteins containing MHC binding peptides

NetMHCIIpan (version 2.1) from the Center for Biological Sequence Analysis (CBS) was used to predict MHC binding peptides for a set of 92 BoLA-DRB3 molecules (Nielsen et al., 2010). NetMHCIIpan by default computes an IC_{50} (nM) peptide-MHC binding affinity score for 15 amino acids at a time sequentially moving one amino acid along the protein (i.e. a sliding window). NetMHCIIpan makes the distinction that an IC_{50} score greater than 50 but less than 500 indicates a weak binding bond; and an IC_{50} score less than 50 a strong bond. Affinity scores to all 92 BoLA-DRB3 alleles were computed. This resulted in thousands of individual peptide-MHC binding scores per protein. Two methods were devised in an attempt to encapsulate the thousands of scores. The first involved counting the number of binding peptides. In effect there were three counts per protein for the total number of weak, strong, and weak or strong binders. The second method involved adding the IC_{50} scores of binding peptides. Similarly, there were three totals per protein for weak, strong, and weak or strong binders. There was a strong positive correlation between the counts or IC_{50} totals and protein length. The counts and IC_{50} totals were therefore divided by the length. A single probability score was determined for each total and count using random forest (a supervised machine learning algorithm) as an indicator that the protein contains appropriate peptides that bind to BoLA-DRB3 alleles (method described in a previous study (Goodswen et al., 2014b). A caveat here is that random forest was trained on proteins expected to be immune-exposed or unexposed proteins and not on known vaccine and non-vaccine proteins.

2.7. Mapping immunological 'hotspots' within a protein

An in-house Perl script was written to implement the following method of mapping immunological 'hotspots'. *C*onsecutive IC_{50} (nM) affinity scores, computed along a protein by NetMHCIIpan (see previous *section 2.6*), were grouped if the score was less than 500. More than one member in a group defined an island of epitopes and the number of members indicated epitope density. High density islands are thought of as immunological 'hotspots'. The grouping was performed for each of the 92 BoLA-DRB3 alleles (i.e. 92 sets per protein were generated defining the number of islands and the number of epitopes per island). The 92 sets were summed and four density characteristics were recorded per protein: total number of islands divided by length of protein, average number of epitopes per island, maximum number of epitopes in the island with the highest density that was specific to one BoLA-DRB3, and the BoLA-DRB3 allele that bound to the highest density island.

2.8. Calculating transcript expression levels

Two separate estimates of the expression level of each NC-Liverpool transcript expressed were generated by the programs RobiNA (Lohse et al., 2012) and Cuffdiff (a part of the Cufflinks package) (Trapnell et al., 2012). The estimates were normalised as RPKM values (reads per kilobase of transcript per million mapped reads). The NC-Liverpool transcripts were those derived from the RNA-Seq experiment described above.

NC-Nowra is considered an attenuated strain in comparison to NC-Liverpool. Live vaccination utilising tachyzoites of NC-Nowra has been shown to prevent abortions when challenged with NC-Liverpool (Williams et al., 2007). As part of the RNA-Seq experiment described above, expression levels of NC-Nowra transcripts were generated by RobiNA. A

295 ‘yes’ or ‘no’ indication of significant differential gene expression between NC-Liverpool and
296 NC-Nowra proteins was recorded. Three methods were used to support the differential gene
297 expression ‘yes’ or ‘no’ indicator: EBSeq (Leng et al., 2013), Cuffdiff (Trapnell et al., 2013)
298 with assembled reference, and Cuffdiff with the original NC-Liverpool reference from
299 ToxoDB.

300

301 ***2.9. Determining protein conservation***

302 The UniProt Reference Clusters (UniRef) (Suzek et al., 2007) provided non-redundant
303 clustered sets of sequences from the UniProtKB that had 100% similarity with a NC-
304 Liverpool protein sequence. Three cluster characteristics were recorded: UniRef cluster
305 name, a list of UniProt IDs and associated organism name for each protein in the cluster, and
306 the total number of proteins in cluster. These characteristics provided clues to the identity of
307 uncharacterised *Neospora* proteins and/or supported or opposed the current annotation. It also
308 highlighted redundant NC-Liverpool proteins.

309 Proteins with at least 50% sequence similarity to NC-Liverpool sequences were also
310 obtained from UniRef. However, the output was filtered to only include proteins originating
311 from apicomplexans. Four cluster characteristics were calculated and recorded: UniRef
312 cluster name; a list containing the total number of proteins originating from specific
313 apicomplexan species groups (the groups are *Babesia*, *Besnoitia*, *Cryptosporidium*,
314 *Cyclospora*, *Eimeria*, *Gregarina*, *Haemoproteus*, *Hammondia*, *Hepatocystis*, *Leucocytozoon*,
315 *Neospora*, *Parahaemoproteus*, *Plasmodium*, *Sarcocystis*, *Theileria*, and *Toxoplasma*); the
316 total number of apicomplexan proteins in cluster; and the total number of proteins in cluster
317 irrespective of its species origin. These characteristics indicated how well an NC-Liverpool
318 protein is conserved among other species and, in particular, apicomplexans. Proteins

containing PAMPs were predicted using the latter and other characteristics (method described later in section 2.12).

2.10. Host similarity

BLASTP was performed between NC-Liverpool and *Bos taurus* protein sequences (downloaded from UniProtKB). The bovine protein associated with the highest bitscore and similarity was extracted from the BLASTP output. Two characteristics were recorded: UniProt ID of the bovine protein with the greatest similarity, and the percentage sequence similarity between the *Neospora* and *Bos* protein.

2.11. Additional information

Additional information on *N. caninum* was extracted from ToxoDB version 12 and recorded: chromosome number encoding the gene of the protein, start and end genomic location of the gene including forward or reverse strand origin, relative location of gene start and end, number of exons, CDS length, molecular weight (a computed value from raw translations that do not take into account any protein or residue modifications), isoelectric point, and the number of transmembrane domains as predicted by TMHMM 2.0. The relative gene start and end was computed in-house and defines the location of the gene start and end position relative to the centre of the chromosome. The relative locations are defined as a percentage e.g. a gene located at the start of chromosome is -100% and indicates that it is located at the furthest distance from chromosome centre i.e. 0% is the centre of the chromosome; and a gene located at end of the chromosome is +100% and also indicates it is the furthest distance from chromosome centre (Fig. 1 and Supplementary data 2).

2.12. Ranking vaccine candidates

There were eight forms of rank values associated with each protein as shown in Table 1. Each rank was computed using the same method but using different protein features selected from Supplementary data S1. The method involved first independently sorting raw values for each selected feature in ascending order (i.e. lowest value at top and largest at bottom of the sorted list) e.g. E_rank has 10 contributing features and therefore 10 independent lists in value ascending order were calculated. The assumption is that the larger the feature value, the greater its importance to candidacy selection (Homolog weight was an exception and ordered in descending order). Furthermore, each feature is assumed to have equal importance in rank determination because the relative accuracy of the feature value is unknown. The next ranking step was to divide the protein position in each ordered list by the number of proteins and then multiply by 100, for example: (1 / 7111) * 100 = 0.01% for protein with lowest value; and (7111 / 7111) * 100 = 100% for highest value protein. Each protein subsequently had a percentage rank for each feature selected (e.g. ten individual percentage ranks for E_rank computation). A single rank was then determined by reordering the proteins based on the magnitude of each individual rank. For example, the protein with the highest E-rank (100%) contained ten rank values that were greater than or equal to the ten rank values of the next highest E-rank, and so on. Note that proteins with exactly the same individual rank values were given the same final single rank.

The W_E_rank is based on the assumption that RNA-Seq evidence is a more reliable indicator of 'protein existence' than other available evidence. Therefore, W_E_rank was computed as per E-rank except only RNA-Seq and Sanger RNA-Seq features were *included* (i.e. only two features used as opposed to ten). However, if proteins were computed with the

same rank then their final ordered positions were determined by the rank of other evidence i.e. a rank as per E-rank except RNA-Seq and Sanger RNA-Seq *excluded.* As a brief example, four proteins (x007, x002, x006, x003) have the ranks 100.0, 100.0, 100.0, 99.9 respectively after an initial ranking using only RNA-Seq evidence i.e. three proteins (x007, x002, x006) have the same rank. The ranks for the four proteins when using other evidence that excludes RNA-Seq are x003 = 100.0, x006 = 99.8, x002 = 99.7, x007 = 99.4. The final rank positions as per W_E_rank for the four proteins would be x006 (highest ranked), x002, x007, and x003 (lowest ranked). That is, only the initial identical ranks were reordered (x003 ignored in his case) based on other evidence ranks and so in effect, RNA-Seq evidence is weighted more than other 'protein existence' evidence.

2.13. Feature selection

Once ranking was done we investigated which features contributed most to the ranking. Consequently, using Supplementary Table S1, which ranks all proteins, we used the Random Forest algorithm to identify the most important features used to predict Final_rank. Random Forest (Breiman, 2001) is a robust machine learning algorithm able to learn the mapping from input features to a target value: either as a classification of a discrete value or regression to a floating point value. In this case, using a subset of Supplementary Table S1 as a training set, we learned the mapping from other features to the floating point Final_rank. One of the advantages of Random Forest compared to other machine learning algorithms is that it is able to infer the relative importance of input features towards the target prediction, assigning a Variable Importance Score, using various measures. This is helpful for interpreting datasets. We used the randomForest package in R, using default parameter

settings. Variable importance is reported using two measures: percentage increase in Mean Square Error and increase in node purity. High values indicate strong variable importance.

Random Forest is unable to deal with categorical features with a large number of values or with text features. So we also excluded the following features from analysis: >90%_Sim, Protein_names, Homolog_name, Homolog_Organism, Homolog_Locus_Tag, Cluster_name_for_50%_similarity, Apicomplexan_member_count_for_50%_similarity, Bovine_UniProtID, Gene_ontology_biological_process, Gene_ontology_molecular_function, Predicted_GO_Function_Term, Gene_ontology_cellular_component, Gene_names, Chromosome, Date_of_last_modification, Subcellular_location, Post_translational_modification, and Function. Clearly the features in S1 most useful for predicting Final_rank are the other ranking features from which they are calculated as they are the most strongly correlated. So, we also excluded the following features from the ranking: E_rank, W_E_rank, V_rank, W_Final_rank, P_rank, Final_P_rank, and W_Final_P_rank.

3. Results

3.1. Ranking of vaccine candidates

Neospora caninum consists of many diverse heterogeneous strains distributed throughout the world (Al-Qassab et al., 2010), but almost all publicly available *N. caninum* protein sequences are from the NC-Liverpool strain. Both the Universal Protein Resource knowledgebase (UniProtKB) and ToxoDB hold similar sets of NC-Liverpool protein sequences. Table 2 shows the extent of *Neospora* protein annotation in UniProtKB. The deduced protein sequences result from predicting genes using various *ab initio* gene predictors supported by expressed sequence tags (ESTs). An NC-Liverpool genome containing 14 pseudo-chromosomes was constructed using supercontigs aligned to 14 publicly available, albeit draft, *Toxoplasma gondii* ME49 chromosomes based on predicted protein sequence similarity (Reid et al., 2012). mRNA sequencing (RNA-Seq) was also used to improve the annotation, but only for genes for which mRNAs were expressed from tachyzoites during experimental laboratory conditions.

Table 3 shows a breakdown of how the *N. caninum* protein sequences were derived. Most sequences have 'predicted' for their 'evidence for existence' annotation in UniProtKB. Furthermore, over 5600 sequences are annotated with ambiguous protein names. Table 4 lists the types of annotated names. This current annotation state raises uncertainty in its reliability given the fact that the majority of sequences remain unverified and uncharacterised. An unknown percentage of invalid or inaccurate sequences will inevitably exist that can lead to erroneously predicted protein characteristics.

Various protein characteristics were obtained from public resources or computationally predicted for every known NC-Liverpool protein and compiled in a Microsoft Excel worksheet (Supplementary data S1). It was important that every protein was included to

avoid introducing preconceived notions of candidates. Considering all proteins is one of the guiding principles of reverse vaccinology. In the Supplementary data S1 there are 79 columns and 7111 rows. Each column contains a specific protein characteristic and each row holds the collection of characteristics per protein. The columns are grouped with coloured headers to denote associated characteristics.

The aim of this study was to select the *N. caninum* vaccine candidates most worthy of laboratory validation as it is unfeasible to validate all proteins. This entailed using the collection of protein characteristics to make an informed selection of candidacy potential. The large number of proteins under consideration made it impractical for such a selection to be a manual process. Hence the ideal strategy sought was automation based on specific selection criteria. However, the first imposing challenge was that there is no one or even a group of protein characteristics that clearly distinguish a vaccine candidate. Consequently, the selection strategy relied, in the current absence of clearly defined target proteins, on exploiting subtle tendencies of known antigen characteristics. The predicament here is that not all proteins with a particular tendency will necessarily be worthy candidates. Conversely, vaccine candidates that do not follow trends will be missed in the selection. As an example, known vaccine candidates have a tendency to be naturally exposed to the immune system, but not all immune-exposed proteins are immunogenic and there are some instances (Tan et al., 2010) of non-exposed proteins observed to induce immune responses. The second major challenge was the unknown reliability of the protein characteristics to the extent that some predicted proteins may not even exist.

In an attempt to address these challenges we ranked and presented all proteins in preference to specifically selecting a subset of vaccine candidates. Fig. 2 shows an overview of the ranking process (a detailed description is in section 2.12). The understanding is that the optimum candidates within the constraints of the protein characteristics are at least identified

among those available. Furthermore, a researcher can select the desired percentage of top ranked proteins for validation centred on laboratory capability and budget. In Supplementary data S1 there are five columns denoting ranking: 1) E_rank – indicates the likelihood that a protein exists and its sequence is correct as compared with other proteins i.e. represents a protein's data reliability potential. For example, 100% indicates the protein with the best data reliability from the list of proteins. All evidence was considered equal when computing E_rank; 2) W_E_rank – similar to E_rank except RNA-Seq evidence is more favourably weighted than any other evidence; 3) V_rank – represents a protein's antigenic potential and the likelihood that a protein will make a more worthy candidate when compared with other proteins in the list; 4) Final_rank – takes into account both E_rank and V_rank. For example, a 100% indicates the most promising vaccine candidate with the most supportive evidence for its existence when compared to all other proteins in the list; and 5) W_Final_rank –similar to Final_rank except W_E_rank and V_rank are used (W_Final_rank defines the current protein order in Supplementary data S1). It is difficult to imply with any great certainty that the top ranked proteins will prove to be worthy vaccine candidates. Nevertheless, higher ranked proteins are considered more likely to be worthy than lower ranked proteins.

To substantiate the ordered list of proteins, other than laboratory validation, we checked the rank of known *Neospora* vaccine candidates from published studies. The majority of known candidates are composed of one or a combination of surface, rhoptry, dense granule, and microneme proteins. These candidates could theoretically be grouped into the 'expected' category based on current knowledge. That is, it is well-documented that an apicomplexan pathogen invades a host cell first by, recognising host-cell surface receptors via antigens on its cell membrane, and then secreting proteins from specialized secretory organelles (rhoptries, micronemes and dense granules) (Chen et al., 2008; Roos, 2005). Table 5 shows a list of known candidates and how they rank. Of the 14 unique proteins, seven were in the top

1% and eight in top 10% (see W_Final_rank column in Supplementary data S1 (sheet 2)). Two of the six proteins not in the top 10% are from the 'expected' category (gene names are ROP2 and MIC1 ranked 81.3% and 80.6% respectively). These proteins were highly ranked as vaccine candidates (i.e. V_rank) but were lowly ranked for their likelihood to exist (E_rank and W_E_rank). This low existence rank was due to either no gene being predicted by any prediction method or no consensus between methods. Notwithstanding the existence rank, two candidates, BAG1 (UniProt ID F0VGW4) and IMP1 (F0V754), were clearly shown not to have vaccine potential. The IMP1 immune mapped protein 1 (IMP1) is an 'unexpected' candidate. In this study IMP1 was computed to be a non-exposed protein (.i.e. its sequence revealed neither a classical signal peptide nor transmembrane region) but has strong evidence of containing peptides that bind to bovine MHC molecules. Interestingly, a study (Cui et al., 2012) demonstrated that IMP1 was found to localize to the membrane of *N. caninum* tachyzoites and speculated that this membrane targeting was instigated by N-myristoylation and palmitoylation. It is unclear how many other *N. caninum* proteins experience similar protein sorting. The protein BAG1 (a small heat shock protein) is expressed during the bradyzoite stage and was computed here as a non-exposed protein. Other bradyzoite stage proteins in the known candidate list are MAG1 (a similar antigen to NcGRA1, NcGRA2, and NcGRA7 (Uchida et al., 2013) and SAG4. Both these antigens were predicted to be exposed with appropriate peptide-MHC binders.

3.2. Pathogen-associated molecular patterns (PAMPS)

Proteins containing PAMPs are considered equally important as antigens in the overall vaccine design strategy. Specific ranks (P_Final_rank or W_P_Final_rank) were assigned to every protein indicating the likelihood they contained PAMPs, supported by existence

evidence, when compared to other proteins. Two types of proteins known to harbour PAMPs are profilin-like (UniProt ID F0V772) and heat shock (F0VMT0). Both were ranked in the top PAMP 5%. Also, protein disulfide isomerase (PDI) is a known candidate and was the top ranked PAMP protein. Interestingly, the known candidates ranked the lowest for vaccine candidacy (V_rank) were all highly ranked PAMP candidates. These candidates were Cyp (cyclophilin), BAG1, and IMP1. The cyclophilin protein has 61% sequence similarity to the bovine cyclophilin protein and therefore maybe involved in molecular mimicry. A known possible mimic, MIF (FOVC39), was nevertheless lowly ranked as a vaccine candidate.

3.3. Feature selection and vaccine selection

Features were investigated for their importance in predicting Final_rank using Random Forest. There was general agreement between the two measures of variable importance (Fig. 3). Most significant features included Exposed, estimating the probability that a protein is exposed to the immune system excluding the MHC binding contributions; Existence, estimating the probability that the protein exists and which incorporates evidence from *ab initio* gene predictors, ESTs and comparative genomics; and Homolog Similarity, which is the percentage sequence similarity between the protein and its closest homolog in the NCBI nr database. Also important were measures of protein abundance (Abundance_1 and Nowra_Abundance) and counts of the number of the peptides in the protein predicted to bind to BoLA-DRB3 (ML_Sum_Count).

Highly ranked proteins (as per V_rank, Final_rank or W_Final_rank) were investigated for their antigenicity potential. Firstly, no close correlation was detected between high

525 epitope density and vaccine candidacy potential. The data and correlation analysis are shown
526 in Supplementary data S3. Secondly, no significant correlation was found between vaccine
527 candidacy potential *and* either the number of homologs, apicomplexan homologs, orthologs,
528 paralogs, and molecular weight, isoelectric point, number of exons, and transcript expression
529 levels. Thirdly, no correlation was detected between gene chromosomal location and vaccine
530 candidacy potential. Figure 1 shows the genomic location for genes from chromosome Ia
531 along with an ortholog gene count (Supplementary data S2 shows similar figures for all
532 chromosomes). Genes that encode highly ranked proteins are distributed throughout the
533 chromosome including the extremities and are equally likely to have large or small ortholog
534 counts.

535

4. Discussion

To address the urgent need for vaccines against a wide range of parasitic diseases including *N. caninum*, we have pursued reverse vaccinology including the development of the tool *Vacceed* (Goodswen et al., 2014c). For many parasitic diseases, it is unknown what will constitute an effective vaccine or the factors providing or contributing to effective protective immunity. Consequently we studied *N. caninum* as an example where reasonable data exists to investigate the application of reverse vaccinology. It is pleasing to note that *Vacceed* has been rapidly adopted by others involved in vaccine development (Palmieri et al., 2017).

We ranked every known publicly available NC-Liverpool protein corresponding to its potential for vaccine candidacy. The top ranked proteins are those that are naturally exposed to the immune system and contain peptide binders to bovine MHC II alleles. This typical target profile was driven by common characteristics of known *N. caninum* candidates. Although strict adherence to such a profile may poorly rank uncharacteristic candidates, it was deemed appropriate for high-throughput ranking in the current absence of clearly defined target proteins. Furthermore, in this study, no significant correlation was found between any compiled protein characteristic and vaccine candidacy potential other than those in the target profile.

To our knowledge, this is the first time that an attempt has been made to identify *N. caninum* candidates using an *in silico* approach. The alternative traditional culture-based approach has so far only identified a few candidates and is conceivably too time-consuming to fulfil the urgency. Whether the top ranked proteins in the list prove to be worthwhile will only be known following challenge trials in cattle. However, this ultimate validation requirement for an *in silico*-derived candidate is not any different to that required for culture-

based derived ones. As a minor endorsement for the *in silico* approach, eight of 14 known candidates were ranked in the top 10% as potential vaccine candidates (11 in top 20%).

The ranked proteins are provided in the Supplementary Table S1 file. This file is a valuable resource for *N. caninum* vaccine researchers as it provides an informed starting point for laboratory testing. The desired percentage of top ranked proteins for validation can be selected according to laboratory capability and budget. Moreover, the different forms of ranks provide options to a researcher during selection that can govern the tolerated number and type of false candidates. For example, selecting highest ranked vaccine candidates (V_rank) without evidence for existence is expected to reduce false negative and increase false positive rates. This is because the majority of NC Liverpool proteins are predicted and unverified. The Final_rank combines V_rank and E_rank in an attempt to balance true and false prediction outcomes. This rank form also has the potential to identify candidates that are expressed only at a specific time point or at undetectable levels i.e. a protein unsupported by RNA-Seq but unanimously predicted by gene predictors. Conversely, the W_Final_rank form unfavourably ranks proteins unsupported by RNA-Seq under the assumption that gene prediction is inaccurate. The rank form ultimately used is at the discretion of the laboratory researcher. Either way, the top ranked proteins are considered the optimum candidates given the current *N. caninum* data, knowledge, bioinformatics programs, and the level of confidence (Final_rank) or uncertainty (W_Final_rank) in gene prediction as perceived by the researcher. Furthermore, the PAMP rank (P_Final_rank or W_P_Final_rank) provides a useful indicator for PAMP-based adjuvants.

The Final_rank in effect was computed from 12 selected features. However, the Random Forest investigation for their importance revealed that only five features were major contributing predictors (Existence, Homolog_Similarity, Abundance_1, Exposed and ML_Sum_Count). The importance of these five features, nevertheless, is specific to

Neospora data. Their importance, and the importance of other features, will likely vary in accordance with feature reliability when applied to data from other species. Nevertheless, the approach described here on feature selection, is applicable more generally to other parasitic diseases.

The expectation is that validation in assays and animal models will initially provide indications of efficacy for highly ranked candidates. These indications can then help refine the search target to re-rank candidates. Fine-tuning of candidate rankings are anticipated following iterative cycles of computer analysis and laboratory validation. Moreover, researchers adopting a similar *in silico* approach to that described in this study can use validation feedback for prediction model refinement. Improving reliability of protein characteristics will ultimately lead to better candidate ranking accuracy.

Acknowledgements

SJG gratefully acknowledges receipt of a PhD scholarship from Zoetis (Pfizer) Animal Health. A special acknowledgment goes to Dr Stephen Bush (UTS, Maths) for his advice on statistics and Stephen Daniels for his work on the parasite cultivation and RNA extraction.

Appendix A. Supplementary data associated with this article can be found, in the online version, at ????

References

Al-Qassab, S.E., Reichel, M.P., Ellis, J.T., 2010. On the biological and genetic diversity in Neospora caninum. Diversity 2, 411-438.

Alaeddine, F., Keller, N., Leepin, A., Hemphill, A., 2005. Reduced infection and protection from clinical signs of cerebral neosporosis in C57BL/6 mice vaccinated with recombinant microneme antigen NcMIC1. Journal of Parasitology 91, 657-665.

Andrianarivo, A.G., Anderson, M.L., Rowe, J.D., Gardner, I.A., Reynolds, J.P., Choromanski, L., Conrad, P.A., 2005. Immune responses during pregnancy in heifers naturally infected with Neospora caninum with and without immunization. Parasitology Research 96, 24-31.

Aosai, F., Chen, M., Kang, H.K., Mun, H.S., Norose, K., Piao, L.X., Kobayashi, M., Takeuchi, O., Akira, S., Yano, A., 2002. Toxoplasma gondii-derived heat shock protein HSP70 functions as a B cell mitogen. Cell Stress & Chaperones 7, 357-364.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H.Z., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.L., 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32, D115-D119.

Banchereau, J., Briere, F., Caux, C., Davoust, J., Lebecque, S., Liu, Y.T., Pulendran, B., Palucka, K., 2000. Immunobiology of dendritic cells. Annual Review of Immunology 18, 767-+.

Borodovsky, M., Rudd, K.E., Koonin, E.V., 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. Nucleic Acids Research 22, 4756-4767.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Brent, M.R., 2007. How does eukaryotic gene prediction work? Nat Biotech 25, 883-885.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Cannas, A., Naguleswaran, A., Muller, N., Eperon, S., Gottstein, B., Hemphill, A., 2003. Vaccination of mice against experimental Neospora caninum infection using NcSAG1- and NcSRS2-based recombinant antigens and DNA vaccines. Parasitology 126, 303-312.

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.

Cho, J.-H., Chung, W.-S., Song, K.-J., Na, B.-K., Kang, S.-W., Song, C.-Y., Kim, T.-S., 2005. Protective efficacy of vaccination with Neospora caninum multiple recombinant antigens against experimental Neospora caninum infection. The Korean journal of parasitology 43, 19-25.

Cui, X., Lei, T., Yang, D.Y., Hao, P., Liu, Q., 2012. Identification and characterization of a novel Neospora caninum immune mapped protein 1. Parasitology 139, 998-1004.

Davies, M.N., Flower, D.R., 2007. Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12, 389-395.

Debache, K., Guionaud, C., Adaeddine, F., Hemphill, A., 2010. Intraperitoneal and intra-nasal vaccination of mice with three distinct recombinant Neospora caninum antigens results in differential effects with regard to protection against experimental challenge with Neospora caninum tachyzoites. Parasitology 137, 229-240.

Debache, K., Guionaud, C., Alaeddine, F., Mevissen, M., Hemphill, A., 2008. Vaccination of mice with recombinant NcROP2 antigen reduces mortality and cerebral infection in mice infected with Neospora caninum tachyzoites. International journal for parasitology 38, 1455-1463.

Del Rio, L., Butcher, B.A., Bennouna, S., Hieny, S., Sher, A., Denkers, E.Y., 2004. Toxoplasma gondii triggers myeloid differentiation factor 88-dependent IL-12 and chemokine ligand 2 (monocyte chemoattractant protein 1) responses using distinct parasite molecules and host receptors. J Immunol 172, 6954-6960.

Donati, C., Rappuoli, R., 2013. Reverse vaccinology in the 21st century: improvements over the original design. Year in Immunology 1285, 115-132.

Dubey, J.P., Schares, G., 2011. Neosporosis in animals-The last five years. Veterinary Parasitology 180, 90-108.

Ellis, J., Miller, C., Quinn, H., Ryce, C., Reichel, M.P., 2008. Evaluation of recombinant proteins of Neospora caninum as vaccine candidates (in a mouse model). Vaccine 26, 5989-5996.
Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nature Protocols 2, 953-971.
Fickett, J.W., 1996. Finding genes by computer: the state of the art. Trends in Genetics 12, 316-320.
Gazzinelli, R.T., Denkers, E.Y., 2006. Protozoan encounters with Toll-like receptor signalling pathways: implications for host parasitism. Nature Reviews Immunology 6, 895-906.
Goodswen, S., Barratt, J., Kennedy, P.J., Ellis, J.T., 2015. Improving the gene-structure annotation of the Apicomplexan parasite Neospora caninum fulfils a vital requirement towards an in silico derived vaccine. International Journal for Parasitology 45, 305-318.
Goodswen, S., Kennedy, P., Ellis, J., 2013a. A novel strategy for classifying the output from an *in silico* vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. BMC Bioinformatics 14, 315.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2013b. A guide to in silico vaccine discovery for eukaryotic pathogens. Brief Bioinform 14, 753-774.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2013c. A review of the infection, genetics, and evolution of Neospora caninum: From the past to the present. Infection, Genetics and Evolution 13, 133-150.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2014a. Discovering a vaccine against neosporosis using computers: is it feasible? Trends Parasitol 30, 401-411.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2014b. Enhancing In Silico Protein-Based Vaccine Discovery for Eukaryotic Pathogens Using Predicted Peptide-MHC Binding and Peptide Conservation Scores. Plos One 9.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2014c. Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. Bioinformatics 30, 2381-2383.
Gross, S., Brent, M., 2006. Using Multiple Alignments to Improve Gene Prediction. Journal of Computational Biology 13, 379-393.
Haldorson, G.J., Mathison, B.A., Wenberg, K., Conrad, P.A., Dubey, J.P., Trees, A.J., Yamane, I., Baszler, T.V., 2005. Immunization with native surface protein NcSRS2 induces a Th2 immune response and reduces congenital Neospora caninum transmission in mice. International journal for parasitology 35, 1407-1415.
Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., Nakai, K., 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Research 35, W585-W587.
Jenkins, M., Parker, C., Tuo, W., Vinyard, B., Dubey, J.P., 2004. Inclusion of CpG adjuvant with plasmid DNA coding for NcGRA7 improves protection against congenital neosporosis. Infection and Immunity 72, 1817-1819.
Jones, D., 2012. Reverse vaccinology on the cusp. Nature Reviews Drug Discovery 11, 175-176.
Kall, L., Krogh, A., Sonnhammer, E.L.L., 2004. A combined transmembrane topology and signal peptide prediction method. Journal of Molecular Biology 338, 1027-1036.
Kawai, T., Akira, S., 2010. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nature Immunology 11, 373-384.
Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Research 12, 656-664.
Kerrigan, A.M., Brown, G.D., 2009. C-type lectins and phagocytosis. Immunobiology 214, 562-575.
Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Journal of Molecular Biology 305, 567-580.
Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M.G., Haag, J.D., Gould, M.N., Stewart, R.M., Kendziorski, C., 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29, 1035-1043.
Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., Usadel, B., 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research 40, W622-W627.
Majoros, W.H., Pertea, M., Salzberg, S.L., 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878-2879.

Monney, T., Hemphill, A., 2014. Vaccines against neosporosis: What can we learn from the past studies? Experimental Parasitology 140, 52-70.
Nielsen, M., Justesen, S., Lund, O., Lundegaard, C., Buus, S., 2010. NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. Immunome research 6, 9-9.
Palmieri, N., Shrestha, A., Ruttkowski, B., Beck, T., Vogl, C., Tomley, F., Blake, D.P., Joachim, A., 2017. The genome of the protozoan parasite Cystoisospora suis and a reverse vaccinology approach to identify vaccine candidates. International journal for parasitology.
Pertea, M., Salzberg, S.L., 2002. Using GlimmerM to Find Genes in Eukaryotic Genomes. John Wiley & Sons, Inc.
Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8, 785-786.
Pinitkiatisakul, S., Friedman, M., Wikman, M., Mattsson, J.G., Lovgren-Bengtsson, K., Stahl, S., Lunden, A., 2007. Immunogenicity and protective effect against murine cerebral neosporosis of recombinant NcSRS2 in different iscom formulations. Vaccine 25, 3658-3668.
Plattner, F., Yarovinsky, F., Romero, S., Didry, D., Carlier, M.-F., Sher, A., Soldati-Favre, D., 2008. Toxoplasma profilin is essential for host cell invasion and TLR11-dependent induction of an interleukin-12 response. Cell Host & Microbe 3, 77-87.
Ramamoorthy, S., Sanakkayala, N., Vemulapalli, R., Jain, N., Lindsay, D.S., Schurig, G.S., Boyle, S.M., Sriranganathan, N., 2007. Prevention of vertical transmission of Neospora caninum in C57BL/6 mice vaccinated with Brucella abortus strain RB51 expressing N-caninum protective antigens. International journal for parasitology 37, 1531-1538.
Reichel, M.P., Alejandra Ayanegui-Alcerreca, M., Gondim, L.F.P., 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.
Reid, A.J., Vermont, S.J., Cotton, J.A., Harris, D., Hill-Cawthorne, G.A., Könen-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.
Roos, D.S., 2005. Themes and variations in apicomplexan parasite biology. Science 309, 72-73.
Rosbottom, A., Gibney, E.H., Guy, C.S., Kipar, A., Smith, R.F., Kaiser, P., Trees, A.J., Williams, D.J.L., 2008. Upregulation of cytokines is detected in the placentas of cattle infected with Neospora caninum and is more marked early in gestation when fetal death is observed. Infection and Immunity 76, 2352-2361.
Seabury, C.M., Seabury, P.M., Decker, J.E., Schnabel, R.D., Taylor, J.F., Womack, J.E., 2010. Diversity and evolution of 11 innate immune genes in Bos taurus taurus and Bos taurus indicus cattle. Proceedings of the National Academy of Sciences of the United States of America 107, 151-156.
Sleator, R.D., 2010. An overview of the current status of eukaryote gene prediction strategies. Gene 461, 1-4.
Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32, W309-W312.
Stanke, M., Tzvetkova, A., Morgenstern, B., 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology 7, S11.
Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282-1288.
Tan, T.G., Mui, E., Cong, H., Witola, W.H., Montpetit, A., Muench, S.P., Sidney, J., Alexander, J., Sette, A., Grigg, M.E., Maewal, A., McLeod, R., 2010. Identification of T-gondii epitopes, adjuvants, and host genetic factors that influence protection of mice and humans. Vaccine 28, 3977-3989.
Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology 31, 46-+.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7, 562-578.

Tuo, W., Zhao, Y., Zhu, D., Jenkins, M.C., 2011. Immunization of female BALB/c mice with Neospora cyclophilin and/or NcSRS2 elicits specific antibody response and prevents against challenge infection by Neospora caninum. Vaccine 29, 2392-2399.

Uchida, M., Nagashima, K., Akatsuka, Y., Murakami, T., Ito, A., Imai, S., Ike, K., 2013. Comparative study of protective activities of Neospora caninum bradyzoite antigens, NcBAG1, NcBSR4, NcMAG1, and NcSAG4, in a mouse model of acute parasitic infection. Parasitology Research 112, 655-663.

van Baren, M.J., Koebbe, B.C., Brent, M.R., 2002. Using N-SCAN or TWINSCAN to Predict Gene Structures in Genomic DNA Sequences, Current Protocols in Bioinformatics. John Wiley & Sons, Inc.

Williams, D.J.L., 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. Infection and Immunity 75, 1343-1348.

Wu, T., Watanabe, C., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859-1875.

Yarovinsky, F., Zhang, D., Andersen, J.F., Bannenberg, G.L., Serhan, C.N., Hayden, M.S., Hieny, S., Sutterwala, F.S., Flavell, R.A., Ghosh, S., Sher, A., 2005. TLR11 activation of dendritic cells by a protozoan profilin-like protein. Science 308, 1626-1629.

Figure and Table legends

Fig. 1. The genomic location on chromosome Ia for parental genes of predicted vaccine and PAMP candidates plotted against an ortholog gene count. PAMP, pathogen-associated molecular patterns. Genomic location for each gene on chromosome Ia is represented by a black unfilled circle and is plotted relative to the centre of chromosome i.e. 0 on x-axis represents the exact centre, -100% is start of the chromosome (5' end) and 100% is end of chromosome (3' end). Chromosome extremities are highlighted by First and Last 10% delineated by vertical red dashed lines. Ortholog count (Y-axis) is the number of genes in different species that evolved from a common ancestral gene by speciation. Two black horizontal lines show the mean (solid line) and median (dashed line) of the ortholog count.

Fig. 2. Overview of considerations made in ranking proteins as potential vaccine candidates.

This schematic shows an overview of the feature collection process. A selection of these protein features constitutes determinants for vaccine candidacy ranking. First, evidence features are gathered to support the existence of a protein as the majority of *Neospora caninum* proteins are predicted. Second, vaccine candidacy potential is predicted based on features encoded within the protein sequence. Features are collected for each of the 7111 proteins available for *Neospora caninum*. The ranking method utilizing selected features is described in detail in section 2.12.

Fig. 3. Analysis by Random Forests of the features used to rank proteins from *Neospora caninum* as vaccine candidates. Variable importance is reported using two measures:

percentage increase in Mean Square Error (%IncMSE) and increase in node purity (IncNodePurity). High values indicate strong variable importance.

Table 1

Features selected for the eight protein ranks considered in this study.

Table 2

Current *Neospora* protein annotation (derived from UniProtKB March 2017).

Table 3

Type of evidence that supports the existence of *Neospora caninum* (NC-Liverpool) proteins (derived from UniProtKB March 2017).

Table 4

Protein names for *Neospora caninum* (NC-Liverpool) proteins (derived from UniProtKB March 2017).

Table 5

Published *Neospora caninum* subunit vaccine candidates and how they rank in the current study.

Table 1

Features selected for the study's eight protein ranks.

Data reliability	Weighted data reliability	Antigenic potential	Combined data reliability and antigenic potential	Combined weighted data reliability and antigenic potential	PAMP potential	Combined data reliability and PAMP potential	Combined weighted data reliability and PAMP potential
E_rank	**W_E_rank**	**V_rank**	**Final_rank**	**W_Final_rank**	**P_rank**	**Final_P_rank**	**W_Final_P_rank**
1. Existence	Existence	Exposed	E_rank	W_E_rank	Exposed	E_rank	W_E_rank
2. RNA-Seq	**RNA-Seq**[a]	ML_Sum_Count	V_rank	V_rank	Api member count	P_rank	P_rank
3. Sanger RNA-Seq	**Sanger RNA-Seq**[a]				Total member count		
4. EST_count	EST_count				Ortholog count		
5. DE_min	DE_min						
6. DE_Max	DE_Max						
7. Abundance_1	Abundance_1						
8. Homolog similarity	Homolog similarity						
9. Homolog Weight	Homolog Weight						
10. Number in cluster	Number in cluster						

[a] **RNA-Seq** and **Sanger RNA-Seq** weighted more in ranking**; Existence** –probability score that protein exists using evidence from *ab initio* gene predictions (AUGUSTUS and GlimmerHMM), ESTs (BLAT and GMAP), comparative genomics (N-SCAN); **RNA-Seq** – probability score that RNA-Seq exons derived from RNA-Seq experiment overlap current *N. caninum* annotated exons e.g. if a known gene contains 6 exons and 5 are overlapped by RNA-Seq exons then probability score = 0.83 (.i.e. 5/6); **Sanger RNA-Seq** – as per RNA-Seq but with Sanger exons; **EST_count** – number of ESTs obtained from the Database of Expressed Sequence Tags (dbEST) overlapping the gene; **DE_min** – minimum differential expression (DE) percentile for orthologous tachyzoite genes differentially expressed between *T. gondii* VEG and NC-Liverpool (Reid et al., 2012). Data obtained from ToxoDB but based on pooled day three and four Sanger RNA-seq experiments with DESeq (Anders and Huber, 2010) computations; **DE_Max** – maximum DE percentile for orthologous tachyzoite genes differentially expressed; **Abundance_1** – RNA-Seq derived abundance approximations for NC-Liverpool transcripts normalised using RPKM as computed by RobiNA (Lohse et al., 2012); **Homolog similarity** – percentage sequence similarity between protein and homolog; **Homolog Weight** – a 0 to 7 score to indicate reliability of the homolog protein name as a source of evidence; **Number in cluster** – number of proteins with 100% similarity in cluster as determined by UniRef; **Exposed** – probability score that protein is exposed to the immune system i.e. membrane-associated or secreted. Score was computed by *Vacceed* but peptide-MHC binding contributions excluded; **ML_Sum_Count** – probability score determined by random forest indicating protein contains appropriate peptides that bind to BoLA-DRB3 92 alleles; **Api member coun**t – total number of apicomplexan proteins with at least 50% sequence similarity as obtained from UniRef; **Total member count** – total number of proteins, irrespective of species, with at least 50% sequence similarity as obtained from UniRef; **Ortholog count** – number of orthologous genes as obtained from OrthoMCL DB version 5.

Table 2
Current *Neospora* protein annotation (derived from UniProtKB March 2017).

Taxonomy ID	Mnemonic	Taxonomy Name	Reviewed[b]	Unreviewed[c]	Total
37089		*Neospora* sp.	0	1	1
761197		*Neospora sp.* A California sea lion	0	1	1
761198		*Neospora sp.* B California sea lion	0	1	1
29176[a]	NEOCA	*Neospora caninum*	6	92	98
572307	NEOCL	*Neospora caninum* (strain Liverpool)	0	10,010	10,010
83675	NEOHU	*Neospora hughesi*	0	7	7

[a] Excludes lower taxonomic ranks; [b] manually annotated by UniProtKB curators; [c] Computer-annotated and not reviewed by UniProtKB curators.

Table 3

Type of evidence that supports the existence of *Neospora caninum* Liverpool proteins (derived from UniProtKB March 2017).

Evidence for existence[a]	Count
Evidence at transcript level	0
Evidence at protein level[b]	1
Inferred from homology[c]	852
Predicted[d]	6258
Total	7111

[a] The 'protein existence' evidence does not give information on the accuracy or correctness of the sequence(s) displayed; [b] Indicates that there is experimental evidence for the existence e.g. partial or complete Edman sequencing, identification by mass spectrometry, X-ray or NMR structure; [c] Indicates that the existence of a protein is probable because clear orthologs exist in closely related species; [d] Used by default if protein is without evidence at protein, transcript or homology level

Table 4

Protein names for *Neospora caninum* Liverpool proteins (derived from UniProtKB March 2017)[a].

Protein name/description	Count
Uncharacterized protein	3387
Putative uncharacterized protein	13
Protein name begins with 'Putative'	846
Protein name contains 'Putative'	493
Protein name contains 'Fragment'	17
Protein name contains 'related'	871
Protein name contains 'Probable'	25
SRS domain-containing protein	225
Protein name contains 'Microneme' or 'MIC'	17
Protein name contains 'Dense granule' or 'GRA'	8
Protein name contains 'Rhoptry' or 'ROP'	30
All other names	1179
Total	7111

[a]The counts are based entirely on the protein name as assigned in UniProtKB and compiled using an in-house Perl script that parsed the Protein names of the 7111 proteins from NC-Liverpool.

Table 5

Published *Neospora caninum* subunit vaccine candidates and how they rank in the current study.

Vaccine candidate[a]	Adjuvant	Protection[d]	Gene name	UniProt ID	Existence Rank[i] (%)	Vaccine Rank[j] (%)	Final Rank[k] (%)	Ref.
Different NcSRS2 iscoms	Different types	Reduced infection	SRS2	F0VIH6[e]	95.32	99.93	99.89	(Pinitkiatisakul et al., 2007)
Recombinant NcMIC1	Ribi	Reduced infection	MIC1	F0VCC5	53.03	96.30	80.64	(Alaeddine et al., 2005)
Native NcSRS2	FIA	Reduction in VT	SRS2	F0VIH6[e]	95.32	99.93	99.89	(Haldorson et al., 2005)
Recombinant NcSRS2 and NcDG1	–	66.7% survival rate	SRS2 DG1	F0VIH6[e] F0VF82	95.32 96.65	99.93 96.65	99.89 99.92	(Cho et al., 2005)
Recombinant strain RB51 expressing *N. caninum* antigen[b]	–	Reduction in VT	MIC1 MIC3 GRA2/6 SRS2	F0VCC5 F0VAA2 F0VLB1[f] F0VIH6	53.03 99.38 96.70 95.32	96.30 93.87 95.36 99.93	80.64 99.87 99.90 99.89	(Ramamoorthy et al., 2007)
Intra-nasal recNcPDI (Protein disulfide-isomerase)	Cholera toxin	90%	PDI	F0VAI6	99.85	88.67	99.68	(Debache et al., 2010)
Recombinant proteins[c]	VSA-3	33%	GRA1 GRA2 MIC10 P24B[h]	F0VF82 F0VLB1 F0VR52	96.65 96.70 99.80	96.65 95.36 96.75	99.92 99.90 99.93	(Ellis et al., 2008)
Recombinant NcROP2	FIA or saponin	Reduced infection	ROP2	F0V7L8	53.73	90.76	81.28	(Debache et al., 2008)

Recombinant NcIMP1 (immune mapped protein 1)	FIA	Inhibited host cell invasion	IMP1	F0V754	81.94	39.17	64.60	(Cui et al., 2012)
Recombinant NcCyP (cyclophilin**)**	ImmuMax-SR and CpG	Immunity in a non-pregnant mouse model	CyP	F0V8G1[g]	95.22	61.33	87.64	(Tuo et al., 2011)
pNcGRA7 (plasmid DNA Coding for NcGRA7)	CpG	Reduced infection	GRA7	F0VF82[h]	96.65	96.65	99.92	(Jenkins et al., 2004)
NcSAG1- and NcSRS2-based recombinant antigens and DNA vaccines	Ribi	Reduced infection	SAG1	F0VIH4	65.43	99.42	90.75	(Cannas et al., 2003)
Recombinant NcBAG1, NcMAG1, or NcSAG4	Oil-in-water emulsion with bitter gourd extract	Reduced infection	BAG1 MAG1 SAG4	F0VGW4 F0VJE8 F0VEM5	85.40 99.37 49.59	36.39 88.75 95.09	61.41 99.71 77.16	(Uchida et al., 2013)

Abbreviations: – = unknown; Ribi = Ribi Adjuvant System; FIA = Freund's incomplete adjuvants; VT = vertical transmission. [a]Vaccine candidates are from studies over the past 10 years that attempt to prevent infection, abortion or vertical transmission; [b]MIC1, MIC3, GRA2, GRA6 and SRS2 antigens were expressed in *Brucella abortus* strain RB51; [c]Four recombinant proteins of *N. caninum* (GRA1, GRA2, MIC10, and p24B), MIC10 and p24B provide best protection; [d]Protection = measured or described protection towards type of challenge, [e]UniProt F0VIH4 is an SRS domain-containing protein (Putative srs29b) but is not a complete sequence; [f]GRA6 sequence is 100% similar to GRA2; P24B is a possible novel protein; [g]Same protein as descrived in publication. The highest ranked *N. caninum* cyclophilin protein is F0VLE9 with 99.3%; [h]100% sequence similarity to H6X1L4 (GRA7); [i]equivalent to W_E_rank (represents a protein's weight data reliability potential); [j]equivalent to V_rank (represents a protein's antigenic potential); [K]equivalent to W_Final_rank (a rank taking into account W_E_rank and V_rank).

Figure 1

Click here to download high resolution image

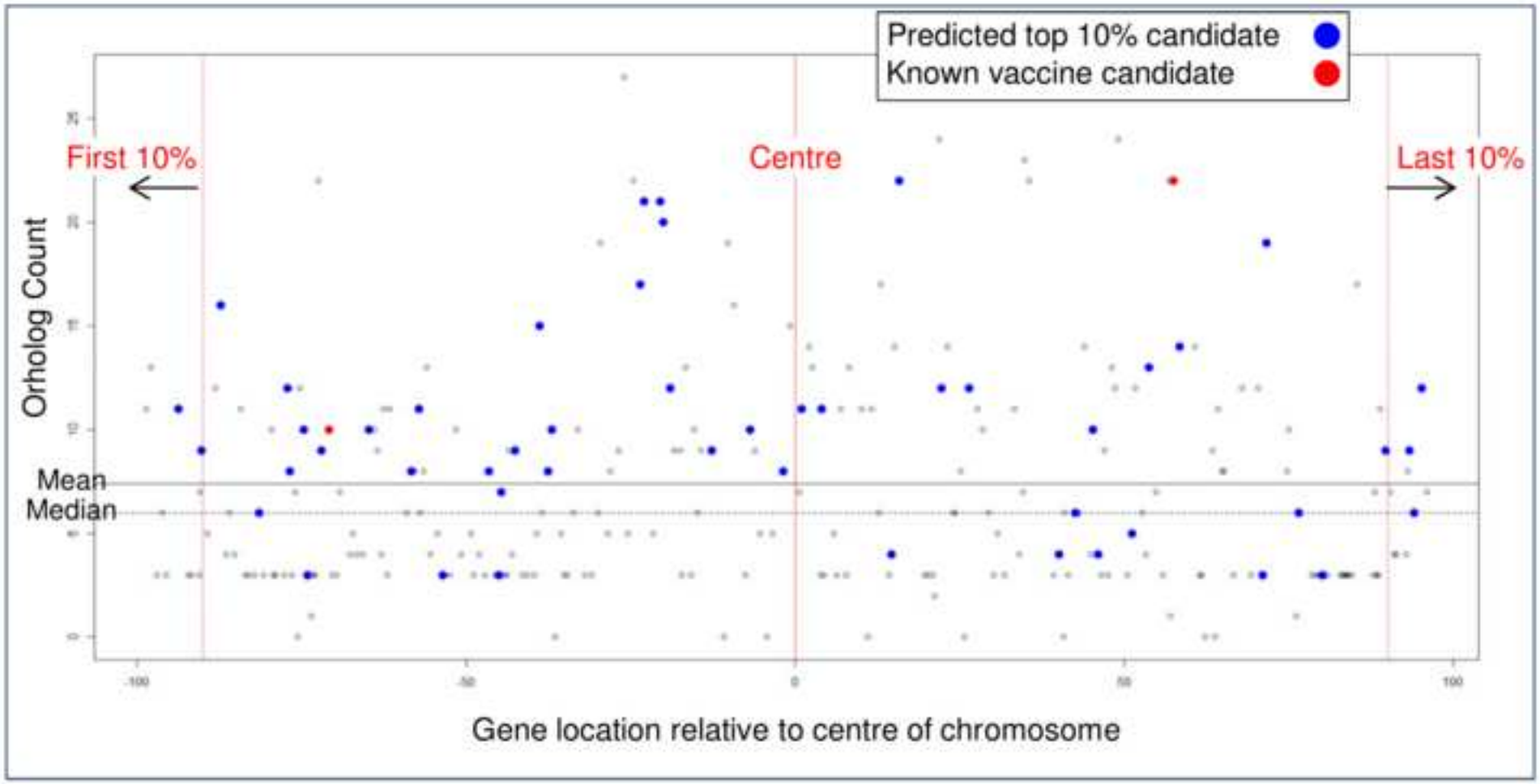


Figure 2

Click here to download high resolution image

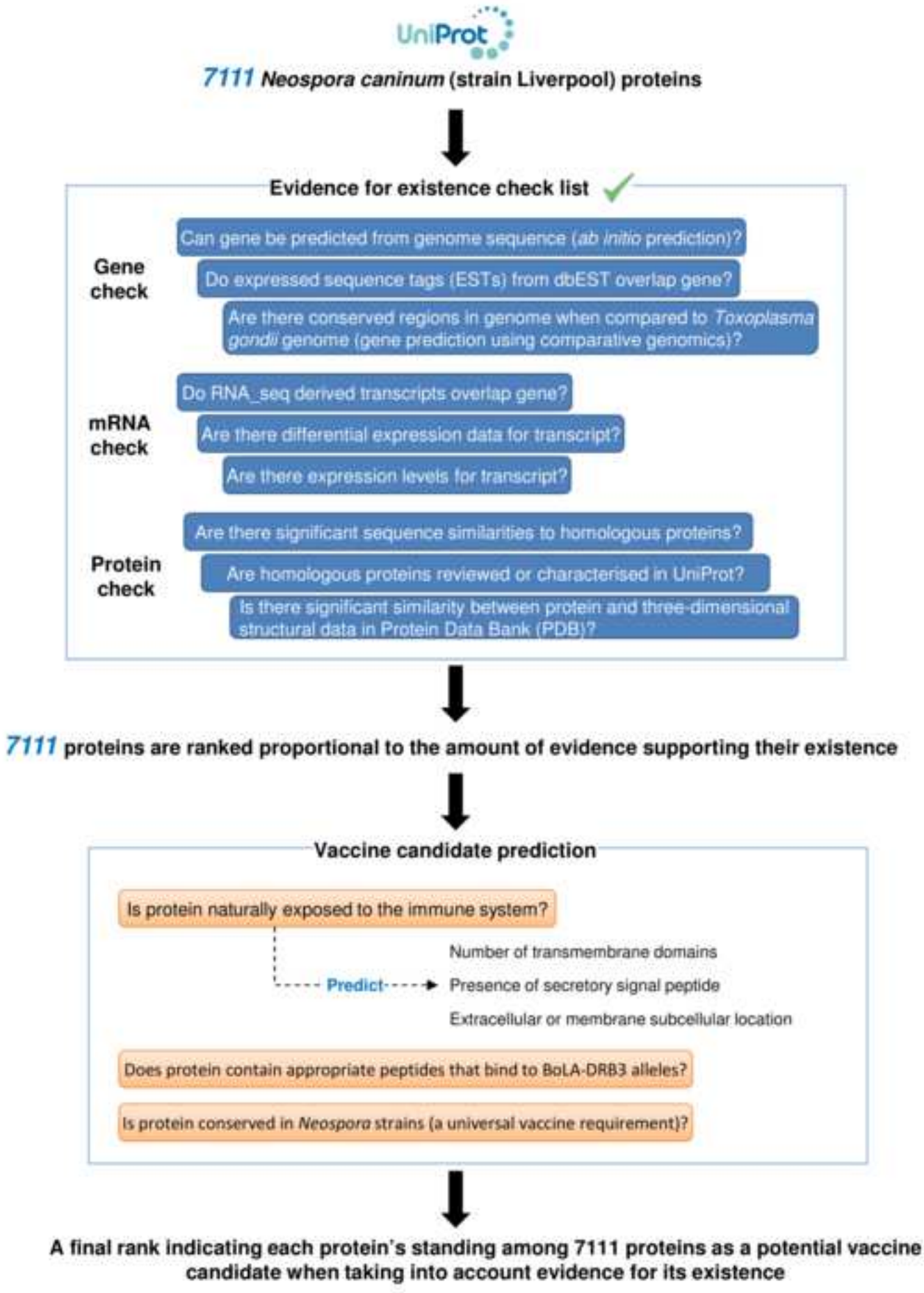
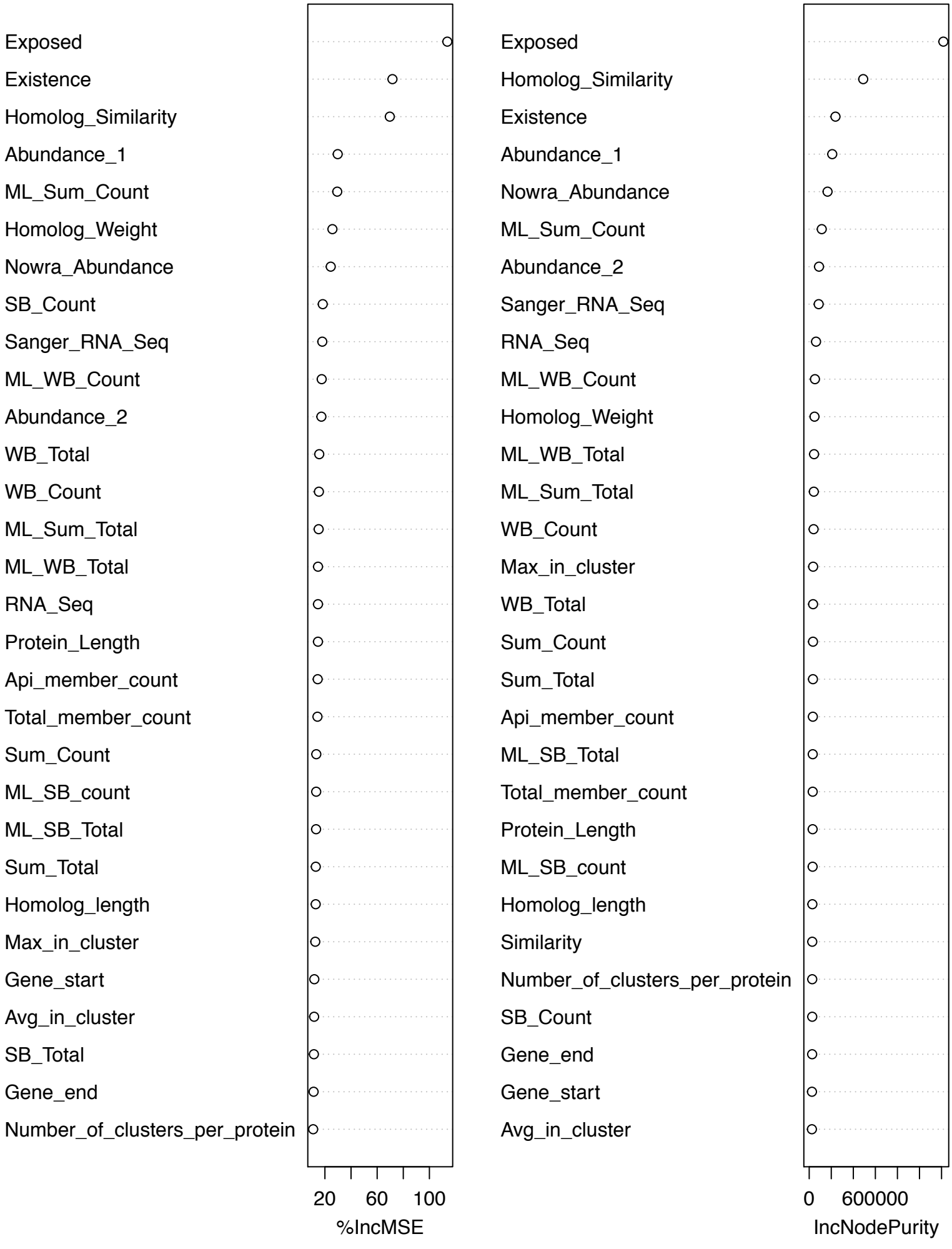


Figure 3


Exposed
Existence
Homolog_Similarity
Abundance_1
ML_Sum_Count
Homolog_Weight
Nowra_Abundance
SB_Count
Sanger_RNA_Seq
ML_WB_Count
Abundance_2
WB_Total
WB_Count
ML_Sum_Total
ML_WB_Total
RNA_Seq
Protein_Length
Api_member_count
Total_member_count
Sum_Count
ML_SB_count
ML_SB_Total
Sum_Total
Homolog_length
Max_in_cluster
Gene_start
Avg_in_cluster
SB_Total
Gene_end
Number_of_clusters_per_protein
20
60
100
%IncMSE
Exposed
Homolog_Similarity
Existence
Abundance_1
Nowra_Abundance
ML_Sum_Count
Abundance_2
Sanger_RNA_Seq
RNA_Seq
ML_WB_Count
Homolog_Weight
ML_WB_Total
ML_Sum_Total
WB_Count
Max_in_cluster
WB_Total
Sum_Count
Sum_Total
Api_member_count
ML_SB_Total
Total_member_count
Protein_Length
ML_SB_count
Homolog_length
Similarity
Number_of_clusters_per_protein
SB_Count
Gene_end
Gene_start
Avg_in_cluster
0
600000
IncNodePurity