SUPPLEMENTAL TEXT

MATERIALS AND METHODS

Genome Assembly

A draft genome was produced using reads from both the Illumina HiSeq and Roche 454 platforms using the following pipeline: first Illumina HiSeq reads were filtered and corrected using the standalone read correction tool ErrorCorrectReads.pl available with ALLPATHS-LG (1). Resulting reads were filtered for possible cross-contamination by aligning reads against human, rhesus macaque monkey and mycoplasma reference genomes using the tool Bowtie2 (2) and removed from subsequent processing steps. Due to their larger read length, Roche 454 single-end reads were filtered for contamination from the same sources using DECONSEQ (3). Reads aligned to the above contaminant genomes with 90% identity over at least 90% of their length were removed from further steps unless they also mapped to *T. gondii*. In addition, for paired-end reads, if either end of the read was aligned under default parameters with any of these possible contaminant genomes, both ends were removed from the data set. Single-end reads were then assembled using the short read assembler Ray (4), using a kmer value of 41 and default parameters. All contigs greater than 500 bp (N50=12,617) were subsequently split into pseudo-reads of length 400 bp, with 200 bp of overlap remaining between adjacent pseudo-reads to maximize support for pre-assembled contig continuity in the next assembly step. Next, after filtering for adapter sequence contamination, 454 reads were combined with the filtered paired-end set of Illumina sequence data and assembled using Newbler v2.8 (454 Life Sciences, Roche Applied Science, Branford, CT). Subsequently, single-end Illumina reads assembled by Ray were added to the assembly project as pseudo-reads in a strategy that has previously been demonstrated in two diverse genome projects (5, 6). 75 contigs under 2000 bp in length were removed from the Newbler assembly as they aligned with the human genome using Bowtie2 with default settings. All 454 reads and assembled pseudo-reads were required to overlap by at least 90 bp with 98% identity. During assembly we ensured that all reads were used only once.

Gene model prediction

Gene models were created using the tools *Genemark-ES*, *Augustus*, *SNAP*, *exonerate*, *EVM, PASA, cufflinks* and *MAKER* (7-14). *Augustus* parameters were obtained from those generally used to predict genes in the closest *S. neurona* relative available, *T. gondii*. *Genemark-ES* was trained automatically using the *S. neurona* scaffolds using default settings. *SNAP* was trained in *MAKER* on the genome of *T. gondii* by running *Augustus,* outputting predictions to zff files, and improving *SNAP* Hidden Markov model (HMM) parameters based on the new predictions. The subsequent HMM was then used to scan the *S. neurona* genome. Expressed sequence tags from *S. neurona* and *T. gondii* were downloaded from NCBI: redundant reads were removed using *CD-HIT* (15) using default parameters. Gene models from *T. gondii, E. tenella,* and *N. caninum* were downloaded from ToxoDB (v8.2) (16). All forms of physical evidence, as well as output from the tools *Augustus* and *Genemark-ES* were then combined using *MAKER*. *SNAP* was then adjusted using the novel gene models, and *MAKER* was run again. Repeats were masked on this second run using *T. gondii*'s library of repeats (available from the Genetic Information Research Institute) as well as repeat families identified in the *S. neurona* genome after 4 rounds of the *ab initio* repeat family prediction tool, *RepeatModeler* (17). In a separate run, *SNAP* was iteratively trained completely independently of the *T. gondii* genome using

MAKER. Gene models were first predicted using the program's est2genome prediction model. Gene models and HMM profiles were then refined seven times through iterative *MAKER* runs using the same physical evidence as above. Output from this final iteration of *SNAP* predictions, as well as the predictions from the self-trained *Genemark-ES* and *Augustus* were then combined using the tool *EVM*. *EVM* also utilized RNAseq evidence aligned against the *S. neurona* genome using *Tophat* and *cufflinks*. Protein and EST alignments were also provided using *exonerate* according to parameters specificied by *EVM.* An assembled alignment of *cufflinks* transcripts and EST transcripts from *S. neurona* was also provided by *PASA* using default parameters.

EVM was found to predict fewer, high-quality genes than *MAKER*, which was found to output more apparently spurious gene models. *EVM* gene models were therefore used in any case that the gene models were predicted in any range overlapping or separate from genes predicted by *MAKER*. *MAKER* annotations are associated with an "edit score" ranging from 0 to 1 and defining the perceived distance between predicted gene models and physical evidence. In cases where *MAKER* predicted gene models in regions of the genome with zero overlap with any *EVM* gene predictions, the gene predictions were retained in the final set of predicted genes only if the AED score was less than 0.1 or when the AED score was below 0.5 and the identified protein was homologous to a *T. gondii* gene (BLASTP e-value less than 1e-20) or contained a PFAM protein domain (18).

Apicoplast sequence and analysis

The apicoplast genome was first sequenced and assembled from *S. neurona* (strain SN3) merozoites maintained in bovine turbinate cell monolayers. Paired-end libraries of 3 and 8 Kbp were generated using the GS FLX titanium rapid library and paired end adaptor kit. Sequences generated from a Roche GS FLX genome sequencer were assembled using Newbler version 2.5.3 (454 Life Sciences, Roche Applied Science, Branford, CT) with the – large parameter turned off. All other parameters were set to default. In order to identify the apicoplast genome, candidate scaffolds were screened by length 25-35 Kbp and tRNA content as identified by tRNAscan-SE (19). We identified a single scaffold of size 34,488 composed of two contigs one of length 5,780 bp and the other 24,002 bp joined by a paired-end read creating a gap of 4,706 bp. Comparison of the two contigs against *Toxoplasma gondii*'s apicoplast genome with BLAST revealed that the smaller contig was part of the inverted repeat (IR) but at roughly half the size. The remaining contig was identified as the other section of the apicoplast genome. The average coverage of the smaller contig (502.8X) was roughly more than twice the coverage of the larger contig (192.5X) indicating the IR had collapsed during assembly. Gene model predictions and annotation was performed using programs *MAKER* (14), and Apollo (20), tRNAscan-SE (18) and ARAGORN (21). *MAKER* was used to scan the apicoplast genome against all apicoplast-encoded protein sequences from *P. falciparum*, *T. gondii* and *E. tenella* obtained from PlasmoDB(22), ToxoDB(23) and Genbank (NC_004823.1) respectively. Output from *MAKER* was loaded into Apollo where coding regions were visualized and aligned to UAA or UAG stop codons and ATG start codons. Hypothetical protein sequences were manually curated to account for alternative codon usage. SSUrRNA and LSUrRNA were found by comparisons to *T. gondii* via BLASTN. tRNAs were found by tRNAscan-SE server using default parameters for a Mito/Chloroplast source (http://lowelab.ucsc.edu/tRNAscan-SE/). The tRNA-Leu in the region after rps4 was found by tRNAscan-SE with search mode: Cove only. tRNAs were further confirmed by the program ARAGORN using default parameters. Mapping of S. neurona SO SN1 reads to the SN3 reference apicoplast sequence was performed with the

GATK pipeline (24).

Gene annotation, ortholog assignment and prediction of syntenic regions

To obtain a prediction of the SRS complement for *S. neurona* gene models were scanned against 8 SRS family domains using HMMER3's *hmmscan* tool (25) and HMM files generated previously (26). HMM alignments were required to be contiguous and over 100 amino acids in length, i-evalue < 0.001 and at least 4 cysteine residues were required over the length of the alignment. Domain families were assigned according to the maximum domain score. To predict orthologs to genes from other apicomplexan species we applied the InParanoid pipeline (27). Metabolic genes were annotated using an inhouse pipeline based on the DETECT enzyme prediction tool (28) as described previously (29). MCSCAN (30) was used to identify syntenic genes, requiring blocks of 3 collinear genes and using an intergenic space of 25,000 bp. Base pair separation was defined as a distance metric. Custom python scripts were then used to export identified links to the genome visualization tool, Circos (31).

Generation and analysis of T. gondii invasion protein co-expression network

For a list of manually curated invasion-related genes for *T. gondii*, we calculated pairwise Pearson correlation coefficients from gene expression data derived from two previously published experiments (32, 33). The first examined gene expression profiles of both tachyzoites and bradyzoites, involving 9 conditions; GEO accession GSE16037. The second was an expression study of the *T. gondii* cell cycle, involving 13 time points - 0 to 12 hours post synchronization; GEO accession GSE19092. A network was then generated in which nodes represent genes and links between nodes indicate significant co-expression (Pearson correlation coefficient > 0.8). Network statistics were calculated using NetworkX (34). The network was visualised using Cytoscape (35).

Homology modeling of SnAMA1a and the SnAMA1a-SnRON2D3 complex

The structural models for *Sn*AMA1a (Ser2 – Ser358; SRCN_465) and *Sn*AMA1b (Ser396 – Cys787; SRCN_461) were generated using Phyre2 (36) in intensive mode based off of a *Tg*AMA1 (PDB ID 2X2Z) model (49 and 44% sequence identity with *Sn*AMA1a and *Sn*AMA1b, respectively), and manually edited in Coot (37). For the complex with *Sn*RON2D3, the region of the *Sn*AMA1a DII loop (His273 – Ala294) corresponding to the region disordered in the *Tg*AMA1 co-structure with a synthetic peptide of *Tg*RON2D3 (*Tg*RON2sp) was removed due to uncertainty in its position while in complex with *Sn*RON2D3. The core 30 residues of *Sn*RON2D3 (His788 to Ile815; SRCN_785) were modelled based on *Tg*RON2sp from the published co-structure with *Tg*AMA1 (38). The *Sn*AMA1a-*Sn*RON2D3 model was refined using Rosetta FlexPepDock (39) with the complex showing the lowest Rosetta energy score chosen and analyzed by visual inspection, PISA (40), ProQ (41), ERRAT (42), and MolProbity (43).

Phylogenetic analyses of ROP kinases

A set of ROP kinase proteins, assigned to individual ROPK subfamilies was obtained for *T. gondii* ME49, *N. caninum* NC1 and *E. tenella* (44). Outliers consisting of shorter sequences and missing key motifs of protein kinase domain were discarded. A global alignment of the kinase domains of these proteins together with the set of *S. neurona* SO SN_1 predicted ROP kinases was then generated using probcons v1.12 (45) and manually edited. TrimAl (46) was used to automatically remove columns of the alignments with a gap threshold of 0.7. A maximum likelihood (ML) phylogeny was reconstructed using PhyML (47) with 1,024 bootstrap replicates using LG substitution model and combined with a Bayesian phylogeny (MrBayes 3.1.2, (48)), using four Markov chains for 4 million generations with a burnin of 25%, using the mixed substitution model and gamma distribution for rate heterogenity. The less paraphyletic MrBayes tree was chosen as the main tree.

Quantitative PCR

Total RNA (2 µg) was isolated from merozoites using the RNeasy mini kit (Qiagen) and reverse transcribed using random primers and SuperScript II (Invitrogen). Gene expression was measured by Taqman qPCR using an Applied Biosystems 7900HT Real-Time PCR System. Primer and probe sets are provided in Table S4 (see supplemental material). The cycling program included 2 min at 50°C, 10 min incubation at 95°C followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. Sarcocystis GAPDH1, ACT1, TUBA1 and EF1 were used as reference genes to normalize the quantity of transcripts (50). Transcript levels were represented as $2^{-\Delta CT}$ to show absolute levels of transcript relative to every gene examined.

Metabolic reconstruction and flux balance analysis

A genome scale metabolic reconstruction for *S. neurona* was initially constructed based on a previous model for *T. gondii* – *i*CS382 (51). This reconstruction was supplemented with enzyme predictions obtained using the DETECT algorithm (28), those supported by both BLAST and PRIAM evidence as we have done previously (39), as well as InParanoid defined orthologs of previously curated *T. gondii* enzymes (27). Non-essential enzymes with gene assignments in *T. gondii* but lacking orthologs in *S. neurona* were removed from the reconstruction. Of the 42 reactions unique to *T. gondii*, 18 were captured in *i*CS382 of which 9 were found to be essential and therefore included in the *S. neurona* reconstruction. The reconstruction is maintained as a spreadsheet (see Table S5 in the supplemental material) in a standard format (51). Flux balance analysis (FBA) was performed using the COBRA Toolbox (version 1.3.4) in MATLAB (52). As previous, each reaction in the model is supplied upper and lower constraints for its flux (for reversible reactions -1000 to 1000 mmol/gDWh; for irreversible reactions 0–1000 mmol/gDWh). In addition, for reactions with single-gene associations, constraints were included in the model based on a previously published set of RNA Seq data for 5 day *T. gondii* (53) as well as our own RNA Seq dataset generated for *S. neurona*. Each reaction receiving a flux constraint based on its associated gene expression relative to the highest gene expression value in the data set and scaled linearly so that the predicted doubling time for *T. gondii* matched the in vivo observation of 11.8 h (54). Single knockouts were simulated for each reaction in the model by setting the constraints of the reaction to 0. Knockout effects were assessed by computing a growth ratio, which is the biomass production rate of the knockout divided by that of the wild type.

Thirteen enzymes encoded in the S. neurona genome previously unreported in T. gondii

We identified 13 enzymes in the *S. neurona* genome which have not previously been reported in *T. gondii*, which all appear expressed in the merozoite stage, include 3-hydroxyacyl-ACP-dehydratase (EC:4.2.1.59) which catalyses an essential step in fatty acid biosynthesis; Very-long-chain 3-oxoacyl-CoA synthase (EC:2.3.1.199) and very-long-chain (3R)-3-hydroxyacyl-ACP-dehydratase (EC:4.2.1.134) which perform steps in fatty acid elongation as well as in the biosynthesis of unsaturated fatty acids; Dihydroorotate dehydrogenase (EC:1.3.5.2) which functions in pyrimidine metabolism (note dihydroorotate oxidase (EC:1.3.98.1) in *T.*

gondii performs a similar role); Glycine dehydrogenase (EC:1.4.4.2), dimethylallyltranstransferase (EC:2.5.1.1) and shikimate kinase (EC:2.7.1.71) which perform critical steps in glycine metabolism, isoprenoid biosynthesis and the shikimate pathway respectively; Phosphoadenylyl-sulfate reductase (EC:1.8.4.8) which is involved in sulphur metabolism; o-succinylbenzoate—CoA ligase (EC:6.2.1.26) and 2-methoxy-6-polyprenyl-1,4-benzoquinol methylase (EC:2.1.1.201), both involved in ubiquinone and other terpenoid-quinone biosynthesis; 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase (EC:2.7.6.3) involved in folate metabolism; farnesyl-diphosphate synthase (EC:2.5.1.10) involved in isoprenoid biosynthesis; and threonine ammonia-lyase (EC:4.3.1.19) involved in serine/threonine metabolism.

The S. neurona genome encodes alpha-glucosidase

An intriguing finding in our annotation efforts was the prediction of an alpha-glucosidase (EC:3.2.1.20) which has the capability of hydrolysing terminal, non-reducing (1->4)-linked alpha-D-glucose residues. While subsequent investigations reveal the presence of orthologs in *T. gondii, N. caninum* and *E. tenella*, we note that there is confusion in their current annotations. For example, the *T. gondii* homolog (TGME49_253030) is annotated in the ToxoDB resource with both alpha-glucosidase and glucan 1,3-alpha-glucosidase (EC:3.2.1.84) activities. Phylogenetic analyses reveal the coccidian homologs to partition with a clade of apparent glucan 1,3-alpha-glucosidases (Figure S5 in the supplemental material). However, within this clade is a protein with experimentally confirmed evidence for alpha-glucosidase activity, while all proteins annotated with glucan 1,3-alpha-glucosidase activity were through less robust electronic annotations. Furthermore, glucan 1,3-alpha-glucosidase activity is performed by a heterodimer in which the alpha subunit performs the catalytic reaction, while the beta subunit is involved in substrate specificity and targeting to the endoplasmic reticulum. While homologs of the alpha subunit were identified in the coccidian and cryptosporidium lineages, we found no evidence for a homolog of the beta subunit in any apicomplexan genome. Further, most of the apicomplexan homologs contain an N-terminal region conserved in parts and interspersed with species-specific insertions and deletions. Together these findings suggest the apicomplexan proteins possess alpha-glucosidase activity, supporting the potential to exploit alternative carbon sources such as maltose, glycogen and sucrose. Since the conversion of sucrose by alpha-glucosidase to fructose and glucose provide additional functionality to our metabolic reconstruction, we performed an *in silico* investigation to examine its potential impact on growth.

Differences with published model (iCS382) of metabolism in T. gondii

New reactions added to iCS382 to simulate T. gondii growth:

Alpha-glucosidase activity was added (EC:3.2.1.20; R00801), and an additional activity for hexokinase was described for the phosphorylation of fructose using ATP (EC:2.7.1.1; R00760). Moreover, fructose and sucrose transport (accompanied by ATP hydrolysis) were added. An artificial diffusion reaction was added for 2-oxobutanoate, originally a deadend metabolite, so as to allow its movement in and out of the system; this reaction was principally added because it was added to the *S. neurona* model.

New reactions added to iCS382 to simulate S. neurona growth:

In addition to the reactions to simulate *T. gondii* growth, two reactions catalysed by EC:4.3.1.19 were added: L-threonine ammonia-lyase (R00996) and L-serine ammonia-lyase activities (R00220). The artificial diffusion reaction for 2-oxobutanoate was added so as to unblock the reaction for L-threonine ammonia-lyase activity which produces 2-oxobutanoate. As per the principle of conservation of mass, if 2-oxobutanoate remained unutilized in the system (therefore a deadend), any reaction producing the metabolite would be blocked, or unusable.

Furthermore, the following nine reactions were retained in the *T. gondii* model, but removed in the *S. neurona* model due to a lack of genetic evidence:

(i) 1.1.1.25
(ii) 1.1.1.31
(iii) 2.4.1.117
(iv) 2.4.1.142
(v) 2.4.2.3
(vi) 2.8.1.2
(vii) 3.5.4.5
(viii) 4.1.3.4
(ix) 6.2.1.16

Implementation of constraints added to the flux balance model

Constraints based on RNA-Seq expression values were applied to the *T. gondii* and *S. neurona* models. Single gene mappings were used to assign constraints to reactions. At the same time, since the construction of the original *T. gondii* model, there have been new releases of ToxoDB. Following reconsideration of original gene mappings, Supplemental Table 2B shows the following changes that affected the application of constraints in the *T. gondii* model.

For further details on the implementation of modeling, please refer to (12).

SUPPLEMENTAL MATERIAL

Text S1. Materials and Methods; Details of *Sarcocystis neurona* metabolic reconstruction; and apicoplast SNP differences.

Table S1 List of ORFs, annotations and expression.

Table S2A. List of *Sarcocystis neurona* IMC proteins. **S2B.** Changes in enzyme:gene assignments for the application of new constraints in *iCS382***.** **S2C.** Predicted impact of single reaction knockouts on parasite growth.

Table S3. List of PCR primers used to confirm SRS expression

Table S4. Metabolic reconstruction for *S. neurona*

Figure S1. Organization of the apicoplast genome sequence and comparison to *Toxoplasma gondii.*
Gene names are as indicated. Red and blue colors indicate the coding strand. Differences with *T. gondii* are indicated on outside of the circle. White circles within genes denote in-frame UGA codons. The *S. neurona* apicoplast genome, like *Toxoplasma* uses an alternate genetic code.

Figure S2. PCR amplification of *rps4* fragment
(A) List of primer combinations used for PCR amplification of the *rps4* fragment insert. (B) Polymerase chain reaction (PCR) was performed using Verbatim high---fidelity DNA polymerase (Thermo Fisher Scientific Inc., Pittsburgh, PA). Three 25 Ml PCR reactions (*S. neurona* SN3 genomic DNA; *S. neurona* apicoplast DNA; or no DNA) were set up for each of the primer pair combinations. The cycling conditions included an initial denaturation at 95^{o}C for 3 min 35 cycles of denaturation at 95^{o}C for 30sec, annealing as shown in the table and extension at 68^{o}C for 1min, followed by a final extension at 68oC for 2min. The amplified PCR products were analyzed on 1.5% agarose gels. The 635bp product amplified using primer pair F1-R3 was purified using a PCR purification kit (Qiagen, Valencia, CA) and was sequenced in both directions at Advanced Genetic Technologies Center, University of Kentucky. (C) Electrophoretic analysis of *rps4* fragment insert PCR reactions PCR product sequence confirms the presence the rps4 fragment insert. Multiple sequence alignment of the sequence from the PCR products, the fragment insert and the original gene show no mutations of the rps4 gene fragment insert from the original gene suggesting it is a relatively recent event (Figure S2).

Figure S3. Alignment of the *rps4* gene and the *rps4* fragment insert
The multiple sequence alignment is composed of 4 sequences: the *rps4* gene, the rps4 fragment insert, and the sequence from each strand of the 635 bp PCR product. Highlighted in yellow is the alignment of the rps4 insert. The single nucleotide highlighted in blue is an ambiguity in the length of the homopolymer run in the *rps4* insert

Figure S4. Conservation of *rpoC2*
(A) Diagram of the *rpoC2* gene across four apicomplexans. (B) Three frame translation of the *rpoC2a* and *rpoC2b* gap. Highlighted in yellow is the stop codon for *rpoC2a*. Highlighted in green is the hypothetical start codon for *rpoC2b*.

Figure S5. Resolving EC annotation of 3.2.1.20 and 3.2.1.84 in apicomplexan homologs
Enzymes with EC 3.2.1.20 and 3.2.1.84 form part of the glycosyl-hydrolase 31 family. The enzyme

corresponding to EC 3.2.1.20 is glucosidase I, a single chain composed of the catalytic subunit. The enzyme corresponding to EC 3.2.1.84 is glucosidase II, a heterodimer composed of the catalytic alpha subunit and the regulatory beta subunit. The multiple sequence alignment of all Swissprot annotated EC 3.2.1.20 and EC 3.2.1.84 sequences (non-redundant at 90% identity) along with apicomplexan homologs (generated using probcons) is shown (right panel), with active site residues indicated by red blocks. The maximum likelihood tree for this MSA generated using PhyML with 1024 bootstrap replicates is also shown (left panel). The evidence used for annotation of the sequences in Swissprot are indicated using coloured asterisks. At the bottom of the figure, the multiple sequence alignment of only the apicomplexan homologs is shown. Black bars correspond to regions with 100% identity, with regions of decreasing greyness corresponding to decreasing identities. The region similar to 3.2.1.20 and alpha subunit of 3.2.1.84 is enclosed in a red box. No homolog of the beta subunit of 3.2.1.84 is found in any of the apicomplexan genomes.

REFERENCES

1. **Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB.** 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences **108:**1513-1518.
2. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:***357*-359.
3. **Schmieder R, Edwards R.** 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS one **6:**e17288.
4. **Boisvert S, Laviolette F, Corbeil J.** 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol **17:**1519- 1533.
5. **Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, Girardo DO, Fodor A, Gusev F, Sanford R, Bruders R, Kittler E, Mills CE, Rast JP, Derelle R, Solovyev VV, Kondrashov FA, Swalla BJ, Sweedler JV, Rogaev EI, Halanych KM, Kohn AB**. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109-114.
6. **Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra MB, Oettler J, Comtesse F, Shih CJ, Wu WJ, Yang CC, Thomas J, Beaudoing E, Pradervand S, Flegel V, Cook ED, Fabbretti R, Stockinger H, Long L, Farmerie WG, Oakey J, Boomsma JJ, Pamilo P, Yi SV, Heinze J, Goodisman MA, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L.** 2011. The genome of the fire ant Solenopsis invicta. Proc Natl Acad Sci U S A **108:***5679*-5684.
7. **Pollier J, Rombauts S, Goossens A.** 2013. Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures. Methods Mol Biol **1011:**305-315.
8. **Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O.** 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:**5654-5666.
9. **Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR.** 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol **9:**R7.
10. **Slater GS, Birney E.** 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6:**31.
11. **Korf I.** 2004. Gene finding in novel genomes. BMC Bioinformatics **5:**59.
12. **Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B.** 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res **34:**W435-439.
13. **Borodovsky M, Lomsadze A.** 2011. Eukaryotic gene prediction using

GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics **Chapter 4:**Unit 4 6 1-10.

14. **Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M.** 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18:**188-196.
15. **Fu L, Niu B, Zhu Z, Wu S, Li W.** 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics **28:**3150-3152.
16. **Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Hu S, Iodice J, Kissinger JC, Kraemer ET, Li W, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ, Jr., Wang H, Warrenfeltz S.** 2013. EuPathDB: the eukaryotic pathogen database. Nucleic Acids Res **41:**D684-691.
17. **Smit AFA, Hubley R** 2008-2010, posting date. RepeatModeler Open-1.0. [Online.]
18. **Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy Sean R, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL.** 2002. The Pfam Protein Families Database. Nucleic Acids Research **30:**276-280.
19. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955-964.
20. **Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME.** 2002. Apollo: a sequence annotation editor. Genome Biol **3:**RESEARCH0082.
21. **Laslett D, Canback B.** 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res ***32:***11-16.
22. **Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Jr., Treatman C, Wang H.** 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res **37:**D539- 543.
23. **Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, Pinney DF, Roos DS, Stoeckert CJ, Jr., Wang H, Brunk BP.** 2008. ToxoDB: an integrated Toxoplasma gondii database resource. Nucleic Acids Res **36:**D553-556.
24. **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.** 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20:**1297-1303.
25. **Eddy SR.** 2011. Accelerated Profile HMM Searches. PLoS Comput Biol **7:**e1002195.
26. **Wasmuth JD, Pszenny V, Haile S, Jansen EM, Gast AT, Sher A, Boyle JP, Boulanger MJ, Parkinson J, Grigg ME.** 2012. Integrated bioinformatic and

targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of Toxoplasma virulence. MBio **3**.

27. **Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL.** 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res **36:**D263-266.
28. **Hung SS, Wasmuth J, Sanford C, Parkinson J.** 2010. DETECT--a density estimation tool for enzyme classification and its application to Plasmodium falciparum. BIOINFORMATICS ***26:***1690-1698.
29. **Hung SS, Parkinson J.** 2011. Post-genomics resources and tools for studying apicomplexan metabolism. Trends Parasitol **27:**131-140.
30. **Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH.** 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res **40:**e49.
31. **Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA.** 2009. Circos: an information aesthetic for comparative genomics. Genome research **19:**1639-1645.
32. **Behnke MS, Radke JB, Smith AT, Sullivan WJ, Jr., White MW.** 2008. The transcription of bradyzoite genes in Toxoplasma gondii is controlled by autonomous promoter elements. Mol Microbiol **68:**1502-1518.
33. **Behnke MS, Wootton JC, Lehmann MM, Radke JB, Lucas O, Nawas J, Sibley LD, White MW.** 2010. Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of Toxoplasma gondii. PLoS One **5:**e12354.
34. **Hagberg AA, Schult DA, Swart PJ.** 2008, p 11-15. 7th Pythin in Science Conference (SciPy2008), Pasadena, CA, USA.
35. **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T.** 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13:**2498-2504.
36. **Kelley LA, Sternberg MJ.** 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc **4:**363-*371.*
37. **Emsley P, Cowtan K.** 2004. Coot: model-building tools for molecular graphics. Acta Crystallographica Section D: Biological Crystallography **60:**2126-2132.
38. **Tonkin ML, Roques M, Lamarque MH, Pugniere M, Douguet D, Crawford J, Lebrun M, Boulanger MJ.** 2011. Host cell invasion by apicomplexan parasites: insights from the co-structure of AMA1 with a RON2 peptide. Science **333:**463-467.
39. **London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O.** 2011. Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. Nucleic Acids Res **39:**W249-253.
40. **Krissinel E, Henrick K.** 2007. Inference of macromolecular assemblies from crystalline state. J Mol Biol **372:**774-*797.*
41. **Wallner B, Elofsson A.** 2003. Can correct protein models be identified? Protein Sci **12:**1073-1086.
42. **Colovos C, Yeates TO.** 1993. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci ***2:***1511-1519.

43. **Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC.** 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr **66:**12-21.

44. **Talevich E, Kannan N.** 2013. Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. BMC evolutionary biology **13:**1-*17.*

45. **Do CB, Mahabhashyam MS, Brudno M, Batzoglou S.** 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome research **15:**330-340.

46. **Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T.** 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. BIOINFORMATICS **25:**1972-*1973.*

47. **Guindon S, Gascuel O.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696-704.

48. **Ronquist F, Huelsenbeck JP.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. BIOINFORMATICS **19:**1572-1574.

49. **Livak KJ, Schmittgen TD.** 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods **25:**402-408.

50. **Song C, Chiasson MA, Nursimulu N, Hung SS, Wasmuth J, Grigg ME, Parkinson J.** 2013. Metabolic reconstruction identifies strain-specific regulation of virulence in Toxoplasma gondii. Mol Syst Biol **9:**708.

51. **Thiele I, Palsson BO.** 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc **5:**93-121.

52. **Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ.** 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat. Protocols *2:*727-738.

53. **Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Konen- Waisman S, Latham SM, Mourier T, Norton R, Quail MA, Sanders M, Shanmugam D, Sohal A, Wasmuth JD, Brunk B, Grigg ME, Howard JC, Parkinson J, Roos DS, Trees AJ, Berriman M, Pain A, Wastling JM.** 2012. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy. PLoS Pathog **8:**e1002567.

54. **Blader IJ, Manger ID, Boothroyd JC.** 2001. Microarray analysis reveals previously unknown changes in Toxoplasma gondii-infected human cells. J Biol Chem **276:**24223-2423