Common Inheritance of Chromosome Ia Associated with Clonal Expansion of *Toxoplasma gondii*

SUPPLEMENTAL MATERIAL

Methods

Parasite harvesting and DNA isolation. Parasites were harvested from human foreskin fibroblasts by passage through 3.0 micron filters (Nucleopore Inc.) and centrifugation at 400 g. For routine genomic DNA isolation, cell pellets were treated with sodium lauryl sulfate, proteinase K (10 µg/ml), 10 mM Tris HCl overnight, extracted with phenol chloroform, and precipitated with ethanol. For large DNA (intact chromosomes) preparation, parasite pellets were resuspended to 1 x 10^8 parasites/ml in 1X PBS and 1% low melting point agarose (Sigma) and, after solidification, the blocks were treated twice in 0.1M EDTA pH8, 0.01 M Tris–HCl, 0.02 M NaCl plus 1% lauroyl sarcosine and 1 mg ml^{-1} proteinase K for 48h at 50°C.

Chromosome Ia and Ib shotgun library preparation. The chromosomal DNA plugs were washed in TE over four hours. They were then loaded on pulsed field gel (5x TB(0.1)E, 1.2% low melting point agarose; Sigma) and run on a CHEF DR II apparatus (Bio-Rad). The power supply was set to 85V, maximum current, a pulse gradient of 1400s-700s , and run for 144 hours. After isolation by pulsed field gel electrophoresis, chromosomal DNA was extracted from low-melting point agarose by loading the gel slice into a 5ml syringe and repeatedly passing the gel through a 23g needle until the gel was a smooth paste. The agarose was digested using beta-agarase (NEB) at 37^{O}C for 30 minutes after which the resulting mixture was extracted with phenol, the aqueous phase recovered and DNA precipitated with ethanol. Recovered DNA was resuspended in 60ul of 1x mung bean buffer, sheared by sonication, then end-repaired by incubation with Mung Bean nuclease (Amersham). DNA fragments were then size-selected by electrophoresis through low-melting point agarose and recovered by digestion with AgarAce (Promega), extraction with phenol and precipitating with ethanol. Libraries with inserts between 2 and 4kb were then constructed by ligating these fragments with pUC19_*Sma*I (Merck).

Annotation of *T. gondii* ChrIa and ChrIb. Annotation was performed using Artemis (Rutherford et al. 2000). On ChrIa and ChrIb, respectively, 211 and 252 putative protein coding sequences

(CDSs) were identified by manual curation of the output of the gene finding software. Gene finding software included Twinscan (Korf et al. 2001), GlimmerM (Salzberg et al. 1999) and SNAP (Korf 2004) trained for *T. gondii*. An average of 5.7 exons per CDS was predicted, with the intron/exon boundaries for some refined using the software 'exonerate' (Slater and Birney 2005). The CDSs are labelled incrementally TgIa.0010, TgIa.0020, etc. Biological functions were ascribed to 155 (33%) CDSs based on manual inspection of the results from FASTA and BLAST sequence similarity searches against the Uniprot protein database and based on protein domain predictions using InterPro (Mulder et al. 2005), TMHMMv2.0 (Krogh et al. 2001), SignalP v3.0 (Bendtsen et al. 2004). Of the functionally annotated CDSs, 4.5% corresponded to previously characterized *T. gondii* genes. Thirteen tRNA genes were predicted using tRNAscan-SE software (Lowe and Eddy 1997). Putative pseudogenes were identified where a region of DNA had significant BLASTX homology to known proteins but possible open reading frames were interrupted by stop codons or frameshifts. The sequence and the annotation described here have been submitted to EMBL with Accession Numbers AM055942 and AM055943.

Alignment to Expressed sequence tags (ESTs) confirmed the expression of 46.5% (219) predicted CDS. The ESTs were obtained from a publicly available database (Li et al. 2004) and from a full length cDNA database (Watanabe 2004).

Tandem repeats ≤ 500bp in length, microsatellites and dinucleotide repeats on ChrIa, ChrIb and a chromosome XII (ChrXII) cosmid were identified using the Tandem Repeats Finder (TRF) program with the default settings (Supplemental Table 1) (Benson 1999).

Comparison of strain RH with strain ME49. We compared the manually finished sequence of ChrIa and ChrIb from the RH strain of *T. gondii* (Berriman 2004) with a whole genome shotgun assembly of the type II ME49 (B7 clone) strain generated by TIGR (TIGR 1999-2002). The sequences of the contigs from ME49 were ordered based on gene synteny using custom-written PERL scripts. BLASTP analysis was used to localize the contigs to their correct region along the finished chromosome sequences with gaps inserted between contigs to simulate possible sequence or physical gaps in the ME49 sequence. The resulting ME49 whole-chromosome scaffolds were

compared with the RH sequences using the diffseq program provided by EMBOSS (Rice et al. 2000) to find differences in one of the following three categories:

1. substitution of one base in ME49 by one in RH (SNP)
2. insertion of one or more bases from RH into ME49
3. replacement of one or more bases in ME49 with one or more bases from RH

The third category was further separated into changes involving tandem repeats (found using TRF (Benson 1999) with the default settings), gene conversions and the remainder. Gene conversions were defined as two regions of the same length differing by two or more bases. The finished sequence of a cosmid from chromosome XII was also compared with the type II strain generated by TIGR. The results of these comparisons are given in Supplemental Table 2.

Identification of genetic markers for ChrIa. We amplified genomic regions flanking SNPs from all three lineages using the type strains: RH (type I), ME49 (type II) and VEG (type III). Amplicons were sequenced in triplicate from separate PCR reactions (sequencing utilized BigDye cycle sequencing (ABI) and was conducted by SeqWright (DNA Technology Services, TX, USA)). The resulting sequences were analyzed using Vector NTI v9 (InforMax, Invitrogen Life Sciences) to assemble consensus sequences for each strain. CLUSTALX was used to align the sequences and identify strain-specific polymorphisms. SNPs were screened for whether they detected RFLPs using NEBcutter v2.0 (Vincze et al. 2003). Markers were analyzed by PCR amplification, restriction enzyme digestion, and agarose gel electrophoresis in the presence of ethidium bromide. Specific information on each of the markers including primers, PCR conditions, RFLPs and allele types can be found online (Khan 2005; Khan et al. 2005).

Coding regions. Comparison of ChrIa and ChrIb revealed the following features: The total coding percentage (including introns) is about 57% for both chromosomes. Gene density (1 gene per 8.5 kb) is lower than that observed in *Plasmodium falciparum* (1 gene per 4.8 kb) (Gardner et al. 2002). ChrIa and ChrIb have 7 and 4 pseudogenes, respectively. This may be attributed to generally larger intergenic regions, more and larger introns. The proportion of spliced genes is similar, 69% for

ChrIa and 75% for ChrIb. This proportion is roughly similar to that of *Theileria annulata* (70%) (Pain et al. 2005) but higher than that for *Plasmodium falciparum* (50%) (Gardner et al. 2002) and *Cryptosporidium* (5%) (Xu et al. 2004). Both chromosomes have an average number of 6 exons per gene. Notably, the average exon length differed, 463 bp for ChrIa and 386 bp for ChrIb, but the average intron lengths are similar at 519 bp and 528 bp respectively. The majority of spliced genes contain 1 to 5 introns (55%) but the number of introns per gene can be as high as 29. The length of the introns also varies widely (from 5 to 1896 bp). The average length of the 5'UTRs is 332 bp and of the 3'UTRs is 371 bp, based on EST evidence.

Gene content. Approximately 58% of the predicted genes did not have sufficient similarity to genes sequenced in other organisms to allow functional assignments. This figure is similar to that reported on completion of the *Plasmodium falciparum* (Gardner et al. 2002) and the *Theileria annulata* genomes (Pain et al. 2005). Expression of 15% of these genes was verified by their representation in the EST dataset. We were able to assign a function to 32% of the predicted genes with another 10% having similarity to uncharacterized genes in other organisms. Of those genes with a functional assignment, 4.2% had been previously identified in *T. gondii* and 76% had similarities to those in other apicomplexans. About 13% of the total predicted proteins have 1 or more transmembrane domain(s) and 24% are predicted to contain a putative signal peptide or signal anchor. A total of 110 Interpro domains were identified. RNA-recognition motif (IPR000504) was the most abundant. Altogether 13 tRNA genes were found. We observed a relatively high number of enzymes on both chromosomes. 46% of the functionally annotated genes on ChrIa and 35% on ChrIb were predicted to encode enzymes. An analysis of the encoded protein sequences with the program SEG (Wootton 1994) shows that the protein-encoding genes are not enriched in low-complexity sequences (15%) to the extent observed in the proteins from Plasmodium (70%).

Gene families and repeats. With the TribeMCL clustering algorithm (Enright et al. 2002) we were able to identify five putative paralogous gene families on the two chromosomes. One family consisting of five members has no characterized function. Four of the five gene families are

clustered and partly dispersed with pseudogenes of the same family, which probably serves to produce diversity. Unlike other apicomplexans, just one gene family maps adjacent to one of the telomeres (Barry et al. 2003) while the other four gene families are localised in non-subtelomeric regions.

Both chromosomes have a comparable number of microsatellites and conventional TTTAGGG telomere repeat units (Supplemental Table 1). No characteristic AT-rich region similar to those reported in *Plasmodium* and *Theileria* were identified. Therefore, a centromere could not be verified. Other repeat sequences occur between the telomere and the subtelomeric region. Altogether 11 different repeat regions have been found, 8 different types on ChrIa and 3 different types on ChrIb. Ten of these repeat sequences have not been previously characterized in *T. gondii*. ChrIa has no repeat families at one subtelomeric end but the other, larger subtelomeric end with a length of 73 kb shows 8 different repeat families. The largest family consists of 14 direct repeat units of 476 bp. ChrIb contains a large repeat family at one subtelomeric end and 2 repeat families at the other. Interestingly, unlike other apicomplexans, the subtelomeric ends of Chrs Ia and Ib are quite distinct. No similarity between the 11 repeat families has been found. Also in comparison with other apicomplexan subtelomeric ends (i.e. *Eimeria*, *Plasmodium*, *Theileria* and *Cryptosporidium)* no similarities between the subtelomeric regions have been observed between the 2 chromosomes.

Supplemental Table 1. Tandem repeats on ChrIa, ChrIb and a ChrXII cosmid

	ChrIa	ChrIb	Chr XII Cosmid
All pattern lengths ≤ 500bp			
Number	336	443	7
Number per kb	0.17	0.22	0.13
Microsatellites (1-6bp)			
Number	185	272	3
Number per kb	0.10	0.14	0.06
Dinucleotide repeats			
Number	159	238	3
Number per kb	0.08	0.12	0.06

Supplemental Table 2. Categories of polymorphism for ChrIa, ChrIb and a ChrXII cosmid

	ChrIa	ChrIb	ChrXII Cosmid
Single nucleotide polymorphisms			
Number	332	12,925	354
Number per kb	0.17	6.42	6.60
Gene conversions*			
Number	21	2251	43
Number per kb	0.01	1.12	0.80
Differences in tandem repeats†			
Number	22	250	2
Number per kb	0.01	0.12	0.04
Other differences‡			
Number	154	981	17
Number per kb	0.08	0.49	0.32

* two regions of the same length differing by two or more bases † includes expansions and contractions with or without substitutions ‡ includes insertions, deletions and substituted regions

Supplemental Table 3. Classification of genes on ChrIa

Function	Start (bp)	Description*	GO terms*
Growth, cell cycle			
	1102408	serine/threonine-protein kinase, putative / NimA-related protein kinase, putative	GO:0004674, GO:0005524, GO:0006468
	182502	zinc finger (CCCH type) protein	
	372136	dyslexia susceptibility 1 candidate	gene 1 homolog, putative
	648557	zinc finger (CCCH type) protein	GO:0003676
Transcription			
	585847	ssl1-like protein	
	1358432	transcription elongation factor s-II	GO:0003677, GO:0003700, GO:0003711, GO:0006354, GO:0006355, GO:0006357
Translation			
	709952	elongation factor 1-alpha, putative	GO:0003746, GO:0005525, GO:0005853, GO:0006412, GO:0006414
	848104	eukaryotic translation initiation factor 3 subunit 8, putative	GO:0003743, GO:0005852, GO:0006446
	800267	eukaryotic translation initiation factor 3, putative	GO:0003743, GO:0006412
	1613874	ribosomal protein l20, putative	GO:0000315, GO:0003723, GO:0003735, GO:0005622, GO:0005739, GO:0005840, GO:0006412
	1214949	yrdC domain protein	
	494401	histidyl-tRNA synthetase-related	
	1454477	translation initiation factor	GO:0003743, GO:0005852, GO:0006446
RNA metabolism			
	544957	ATP-dependent RNA helicase, putative	GO:0003676, GO:0004004, GO:0004386, GO:0005524, GO:0008026
	475458	ATP-dependent RNA helicase, putative	GO:0003676, GO:0004004, GO:0004386, GO:0005524
Nucleic acid biosynthesis			
	1401494	dihydroorotase	GO:0004151, GO:0006207
	46953	nucleoside diphosphate kinase, putative	GO:0004550, GO:0005524, GO:0006183, GO:0006228, GO:000624
	825461	ribonucleoside-diphosphate reductase, large chain	GO:0004748, GO:0005971, GO:0006260
Replication, recombination, repair			
	1513395	DNA ligase IV, putative	GO:0003910, GO:0005524, GO:0006260, GO:0006281, GO:0006310
	40392	DNA repair protein rad23 homolog b, putative	GO:0003685, GO:0006289
	600576	N-glycosylase/DNA lyase-related	
Histone modification			
	1564179	histone acetyltransferase, putative	
	1471827	poly(ADP)-ribose polymerase-related	

Function	Start (bp)	Description*	GO terms*
Protein modification			
	931885	methyl transferase, putative	
	1570076	arginine N-methyltransferase-related	GO:0016273
	248879	ubiquinone/menaquinone biosynthesis methyltransferae	GO:0005739, GO:0030580, GO:0045426
	1174926	tRNA (guanine-N(7)-)-methyltransferase, putative	GO:0008168, GO:0016020
	1174926	tRNA (guanine-N(7)-)-methyltransferase, putative	GO:0008168, GO:0016020
Ubiquitination			
	1205898	zinc finger, C3HC4 type (RING finger) domain protein	
	255150	UBA/TS-N domain (ubiquitin-associated domain)-containing protein	
	1078967	ubiquitin-protein ligase 1, putative	GO:0004842, GO:0005622, GO:0006464, GO:0006511, GO:0006512
	210257	ubiquitin-protein ligase	GO:0004842, GO:0006512
	1010610	ubiquitin carboxyl-terminal hydrolase, putative	GO:0004197, GO:0004221, GO:0006511
Protein degradation			
	1074011	putative Der1-like protein	GO:0016020
	195638	endopeptidase, putative	GO:0004245, GO:0006508
	1273140	calpain-7, putative	GO:0004198, GO:0005622, GO:0006508
	365781	ATP-dependent Clp protease adaptor protein C, putative	
	1074011	putative Der1-like protein	GO:0016020
Cell signalling			
	891950	mitogen-activated protein kinase-related	GO:0004672, GO:0005524
	1320889	DHHC zinc finger domain-containing protein	GO:0005515
	1703630	inositol phosphatase, putative	
	540274	patched family protein, fragment	
	423284	phosphatidylinositol 4-kinase, putative	GO:0004430, GO:0007165, GO:0016310, GO:0016773
	789529	rhomboid-like protease 5	GO:0016020
	157922	SET-domain protein, putative	GO:0016020
	891950	mitogen-activated protein kinase-related	GO:0004672, GO:0005524
	1770832	3', 5'-cyclic nucleotide phosphodiesterase domain-containing protein	GO:0004114, GO:0004117, GO:0007165
Calcium related			
	1311600	EF hand domain-containing protein	GO:0005509
	321690	calcium-dependent protein kinase	GO:0004674, GO:0005509, GO:0006468
	1273140	calpain-7, putative	GO:0004198, GO:0005622, GO:0006508
	1736982	spry domain containing protein	GO:0016020
Nuclear trafficking			
	502640	importin beta-1 subunit, putative	GO:0005648, GO:0006607, GO:0008139
	1588150	ran-binding protein, putative	GO:0003929, GO:0005525, GO:0006260

Function	Start (bp)	Description*	GO terms*
Intracellular trafficking			
	113096	CGI-141 protein homolog, putative	
	895850	dynein heavy chain, putative	GO:0003777, GO:0007018, GO:0008567
	1430312	translocation protein sec62, putative	GO:0006614, GO:0008565, GO:0015031, GO:0016020, GO:0016021
	549223	vacuolar protein sorting protein	
	1536634	variably charged protein X-C-related / VCX-C protein-related	GO:0003774, GO:0007018, GO:0030286
Metabolic			
	1154466	glucose-6-phosphate-1-dehydrogenase	GO:0004345, GO:0006006, GO:0006010, GO:0006098, GO:0017057
	1715265	NADP-specific glutamate dehydrogenase, putative	GO:0004354, GO:0006536
	1286591	PEP phosphonomutase, putative	
	292052	sulfite oxidase, putative	GO:0006118
	1508876	molybdopterin biosynthesis MoeA protein, putative	
Cellular homeostasis			
	1833437	heat shock protein 90, putative	GO:0003754, GO:0003773, GO:0005524, GO:0005739, GO:0006457, GO:0009408, GO:0016887
	1238989	thioredoxin, putative	GO:0005489, GO:0006118, GO:0006979, GO:0030508
	8556	tubulin-specific chaperone a, putative	GO:0007022, GO:0017072
Mitochondrial			
	1556094	cytochrome c-type heme lyase	GO:0004408, GO:0005739, GO:0005743, GO:0006118, GO:0018063
	1583519	mitochondrial carrier protein	GO:0005488, GO:0006810
	578886	leucine rich repeat protein, putative	GO:0005739
Multiprotein complex			
	1615963	WW domain containing protein	
	298768	TPR domain-containing protein	
Immune related			
	1346791	inflammatory profilin	
	1529063	protein phosphatase 2C, putative	GO:0003824, GO:0004722, GO:0006468, GO:0008287
T. gondii related			
	153760	bradyzoite surface antigen BSR4-related	
	445034	rhoptry protein 4	
	1051573	EGF-like domain-containing protein	GO:0005488, GO:0016020, GO:0020007
	726372	armadillo/beta-catenin-like repeat-containing protein	GO:0005488, GO:0016020, GO:0020007
	662441	type I fatty acid synthase, putative	GO:0003824, GO:0004314, GO:0006633, GO:0008146, GO:0020011, GO:0048037

* The classification of the genes is from our annotation supplemented by information from ToxoDB (http://www.toxodb.org/)

SUPPLEMENTAL MATERIAL REFERENCES

Barry, J.D., M.L. Ginger, P. Burton, and R. McCulloch. 2003. Why are parasite contingency genes often associated with telomeres? *International Journal for Parasitology* **33:** 29-45.

Bendtsen, J.D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology* **340:** 783-795.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27:** 573-580.

Berriman, M. 2004. The *Toxoplasma gondii* sequencing project, pp. http://www.sanger.ac.uk/Projects/T_gondii/. The Wellcome Trust Sanger Institute, Hinxton.

Enright, A.J., S. Van Dongen, and C.A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30:** 1575-1584.

Gardner, M.J., N. Hall, E. Fung, O. White, M. Berriman, R.W. Hyman, J.M. Carlton, A. Pain, K.E. Nelson, S. Bowman, I.T. Paulsen, K. James, J.A. Eisen, K. Rutherford, S.L. Salzberg, A. Craig, S. Kyes, M.-S. Chan, V. Nene, S.J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M.W. Mather, A.B. Vaidya, D.M.A. Martin, A.H. Fairlamb, M.J. Fraunholz, D.S. Roos, S.A. Ralph, G.I. McFadden, L.M. Cummings, G.M. Subramanian, C. Mungall, J.C. Venter, D.J. Carucci, S.L. Hoffman, C. Newbold, R.W. Davis, C.M. Fraser, and B. Barrell. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498-511.

Khan, A. 2005. Alignments of sequenced regions for ChrIa from different strains of *Toxoplasma*, pp. http://toxomap.wustl.edu/chrIa.

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

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59.

Korf, I., P. Flicek, D. Duan, and M.R. Brent. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** S140-148.

Krogh, A., B. Larsson, G. von Heijne, and E.L.L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **305:** 567-580.

Li, L., J. Crabtree, S. Fischer, D. Pinney, C.J. Stoeckert, Jr., L.D. Sibley, and D.S. Roos. 2004. ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Res* **32 Database issue:** D326-328.

Lowe, T.M. and S.R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25:** 955-964.

Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A.N. Nikolskaya, S. Orchard, M. Pagni, C.P. Ponting, E. Quevillon, J. Selengut, C.J.A. Sigrist, V. Silventoinen, D.J. Studholme, R. Vaughan, and C.H. Wu. 2005. InterPro, progress and status in 2005. *Nucleic Acids Research* **33:** D201-205.

Pain, A., H. Renauld, M. Berriman, L. Murphy, C.A. Yeats, W. Weir, A. Kerhornou, M. Aslett, R. Bishop, C. Bouchier, M. Cochet, R.M.R. Coulson, A. Cronin, E.P. de Villiers, A. Fraser, N. Fosker, M. Gardner, A. Goble, S. Griffiths-Jones, D.E. Harris, F. Katzer, N. Larke, A. Lord, P. Maser, S. McKellar, P. Mooney, F. Morton, V. Nene, S. O'Neil, C. Price, M.A. Quail, E. Rabbinowitsch, N.D. Rawlings, S. Rutter, D. Saunders, K. Seeger, T. Shah, R. Squares, S. Squares, A. Tivey, A.R. Walker, J. Woodward, D.A.E. Dobbelaere, G. Langsley, M.-A. Rajandream, D. McKeever, B. Shiels, A. Tait, B. Barrell, and N. Hall. 2005. Genome of the Host-Cell Transforming Parasite *Theileria annulata* Compared with *T. parva*. *Science* **309:** 131-133.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16:** 276-277.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944-945.

Salzberg, S.L., M. Pertea, A.L. Delcher, M.J. Gardner, and H. Tettelin. 1999. Interpolated Markov Models for Eukaryotic Gene Finding. *Genomics* **59:** 24-31.

Slater, G. and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31.

TIGR. 1999-2002. *Toxoplasma gondii* Genome Project, pp. http://www.tigr.org/tdb/e2k1/tga1/. The Institute for Genomic Research, Rockville.

Vincze, T., J. Posfai, and R.J. Roberts. 2003. NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic Acids Research* **31:** 3688-3691.

Watanabe, J. 2004. Full Toxoplasma, pp. http://fullmal.hgc.jp/tg/index.html.

Wootton, J.C. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computational Chemistry* **18:** 269-285.

Xu, P., G. Widmer, Y. Wang, L.S. Ozaki, J.M. Alves, M.G. Serrano, D. Puiu, P. Manque, D. Akiyoshi, A.J. Mackey, W.R. Pearson, P.H. Dear, A.T. Bankier, D.L. Peterson, M.S. Abrahamsen, V. Kapur, S. Tzipori, and G.A. Buck. 2004. The genome of *Cryptosporidium hominis*. *Nature* **431:** 1107-1112.