How old are the extant lineages of *Toxoplasma gondii*?

David A. Morrison

Department of Parasitology (SWEPAR), National Veterinary Institute, 751 89 Uppsala, Sweden.

Running head: Age of *T. gondii* lineages

Correspondence:

Dr David Morrison

Department of Parasitology (SWEPAR),

National Veterinary Institute,

751 89 Uppsala, Sweden

Tel.: +46-18-674161

Fax: +46-18-309162

E-mail: David.Morrison@bvf.slu.se

Abstract. Most known isolates of *Toxoplasma gondii* belong to one of only three lineages, which are presumed to be clonal. Three models have been proposed for the evolutionary relationship of these lineages to the other extant lineages: Model (a) proposing that all lineages are derived from a most recent common ancestor (MRCA) in the distant past, Model (b) that all lineages are derived from a MRCA in the very recent past, and Model (c) that the clonal lineages share a recent MRCA but are related to the other lineages only in the distant past. Here, I test these models using DNA intron and coding-sequence data for loci at 14 genes, using three different methods to calculate the time of the MRCA. All of the calculations agree that the MRCA of the clonal lineages was >70% of the age of the MRCA of all lineages, thus favouring Model (a). The MRCA may have existed ~150,000 years ago, with the clonal lineages expanding in prevalence ~10,000 years ago.

Key words: *Toxoplasma gondii*, most recent common ancestor, genetic distance, coalescent, patristic distance.

Introduction

Toxoplasma gondii (Apicomplexa: Coccidia: Sarcocystidae: Toxoplasmatinae) is a widespread intracellular parasite that uses felids as its definitive host and most homoeothermic species as its intermediate host. It thus has high prevalence in human-associated species, including humans, farm animals and companion animals (Sibley, 2003). The species has an unusual genetic structure (Grigg *et al.*, 2001; Su *et al.*, 2003), with most (>90%) of the known isolates belonging to one of only three lineages (types I, II and III). Lineage II is the most prevalent worldwide, with >75% of the strains, but lineage I is the most virulent in mice. These lineages are genetically almost identical and are thus assumed to be clonal, propagating principally without recourse to the definitive host (Johnson, 1997). The three lineages are genetically very similar, consisting of different combinations of only two alleles, suggesting that they are the product of two parental gene pools.

Thus, the current *T. gondii* pandemic consists of only a few "successful" lineages, comprising those that are in some way adapted to human-associated species, and which are assumed to have become more prevalent with the relatively recent expansion of the populations of humans, agricultural animals and companion animals. However, the origin and evolutionary history of this unusual population structure are still the subject of debate. Understanding the evolution of the lineages should provide a foundation for understanding the genetic variability within the species, which in turn will affect the development of parasite-control strategies.

Three models have been proposed for the evolutionary history of the various extant lineages of *T. gondii* (Fig. 1). These models have much in common, including the suggestion that many of the lineages have involved meiotic recombination, particularly between the clonal lineages and the rarer (here called "atypical" but often called "exotic" elsewhere) lineages, and that the time of the most recent common ancestor (TMRCA) of the extant

lineages is more recent than the origin of the species itself, due to a genetic bottleneck (Grigg *et al*., 2001). Where the models differ is in the proposed time of the most recent common ancestor for the various lineages. Model (a) (Grigg and Suzuki, 2003) posits that all of the lineages are derived from a most recent common ancestor (MRCA) in the distant past, long before the recent population expansion of the clonal lineages. Model (b) (Grigg and Suzuki, 2003) proposes that all of the lineages are derived from a MRCA in the recent past, at approximately the same time as the population expansion of the clonal lineages. Model (c) (Su *et al*., 2003) is somewhat of a hybrid between the other two, with the clonal lineages having a MRCA in the recent past and sharing a MRCA with the atypical lineages only in the distant past. In the absence of a fossil record, Su *et al*. (2003) examined these models by evolutionary analysis of a broad compilation of DNA sequence data, concluding that the TMRCA for all of the lineages was 10^6 years ago while the TMRCA for the clonal lineages was 10^4 years ago, thus favouring Model (c).

However, Su *et al*. (2003) misattribute one of their time estimates for the evolutionary events, thus producing an incorrect TMRCA for the clonal lineages. Here, I correct the misattribution of Su *et al*. (2003) and thus provide an improved estimate of the TMRCA of the clonal lineages. I then address the question of this TMRCA in more detail, examining whether the estimate is robust to the form of the data analysis. I show that the time derived from the method used by Su *et al*. (2003) is probably an underestimate, and that the current sequence data firmly support Model (a).

Materials and methods

I constructed a dataset of aligned nucleotide sequences following the general protocol of Su *et al*. (2003), but of a larger size. Data were compiled from the available literature sources (Table 1), and then supplemented with all of the appropriate sequences available in the

DDBJ/EMBL/GenBank databases. Gene loci were included for analysis if there were sequences for at least two of the three clonal lineages as well as two other sequences. The final dataset comprised >270 kilobases from a total of 20 loci. The main data used in the analyses consisted of introns of seven housekeeping genes and the coding regions or introns of seven antigen genes (Table 1). Each sequence was classified into one of four *T. gondii* groups, based on literature references (e.g. Ajzenberg *et al.*, 2002, 2004): clonal lineage I, clonal lineage II, clonal lineage III, atypical (i.e. the rarer strains). Strains were only included for analysis if they could unambiguously be assigned to one of these classes. TMRCA comparisons were made pairwise between each of the three clonal lineages (i.e. three comparisons) and between each clonal lineage and the atypical group (i.e. three comparisons).

The way in which the data were compiled differs somewhat from what has traditionally been done when studying *T. gondii*. Notably, most authors have used solely their own sequence data, explicitly rejecting the database sequences because those sequences often conflicted with their own, particularly with regard to the clonal lineages. These authors adopted this strategy because of doubts about the unknown quality of the database sequences. On the other hand, I have included every available sequence that could be accurately aligned, irrespective of its provenance. I have done this because it would be a circular argument to insist that *T. gondii* consists predominantly of clonal lineages when this claim is based on data that have been sanitized of all sequence conflicts among the representatives of the clonal lineages (i.e. the evidence for clonality would be artefactual). However, I recognize that my approach runs the counter risk that some of the apparent nucleotide differences among the sequences are laboratory artefacts, either in the laboratory stocks (Howe and Sibley, 1994) or during the sequencing itself (Barry *et al.*, 2003). This prospective problem is real, because it is obvious that for many of the loci there are sequence positions where there is more similarity among different laboratory strains when sequenced in the same laboratory than there is within the same strain when sequenced in different laboratories. I have therefore

chosen what I consider to be the lesser of two evils.

Genetic distances between the sequences were calculated using the PAUP* v4.0b10 program (Swofford, 2002). All pairwise distances between sequences were calculated using the HKY+invariant model of nucleotide substitution, which was the simplest model that could realistically be applied individually to each of the gene loci. The proportion of invariant sites was set to 0.80 for each gene. The pairwise distances were calculated separately for each gene, and then averaged across all of the sequences for each gene for each of the comparisons between the four classes. The overall average across genes was then weighted using the sequence length of each gene.

Coalescence times for groups of sequences were calculated using the IM v11/10/2004 program (Hey and Nielsen, 2004). The four classes of *T. gondii* sequences were compared pairwise using a simple isolation model (i.e. no migration was allowed). The HKY model of nucleotide substitution was used for each gene, with the upper bounds on the priors set to 10 for each theta and the TMRCA. Each analysis consisted of two independent chains started from different seeds, which were then averaged. Each chain consisted of a burnin of 100,000 steps followed by 3,000,000 sample steps, unless mixing was poor, when 500,000 and 10,000,000 steps were used, respectively. The average distance for the comparisons between classes was then calculated. Multilocus analyses were performed separately for the housekeeping and antigen-coding genes — it is important to use multilocus analyses when estimating TMRCAs because the coalescence time of a set of sequences from any one gene does not have to be the same as divergence time of the populations/species (Edwards and Beerli, 2000).

Patristic distances to the MRCA were calculated using the PAUP* v4.0b10 program (Swofford, 2002). A maximum-likelihood tree of all of the sequences was produced separately for each gene, using the HKY+invariant model of nucleotide substitution (proportion of invariant sites = 0.80, empirical base frequencies), the UPGMA starting tree,

tree-bisection-reconnection branch swapping, and with the molecular clock enforced. The TMRCA distance for each of the *T. gondii* classes was defined as the greatest pathlength distance between any of the component sequences and the MRCA in question. The overall average across genes was then weighted using the sequence length of each gene. As an alternative strategy, the gene sequences for each strain were first concatenated. A consensus sequence could then be produced for each of the three clonal lineages, but only three of the atypical strains (CASTELLS, COUGAR, MAS) could be included due to lack of data (i.e. most strains have not been sequenced for most of the loci), and the SAG1 locus was excluded because it has not been sequenced for any of these three strains. After testing for recombination (see below), the CASTELLS sequence was removed as probably being a recombinant. A single maximum-likelihood tree was then produced for the remaining data, as above.

Equality of substitution rates along the evolutionary branches was tested by likelihood-ratio tests (Posada, 2003). All of the maximum-likelihood trees were re-estimated (with the HKY+invariant model) both with and without the branch lengths constrained to be clock-like. The likelihoods from these analyses were then used to perform the likelihood-ratio tests, testing whether the clock-like tree is any worse a fit to the data than is the unconstrained tree. The degrees of freedom for the chi-square approximation were set equal to the difference between the number of nodes and the number of branches, sine the trees contained multifurcations.

Recombination was tested using the RDP v2beta08 program (Martin et al., 2005) and the RecombiTEST suite (Piganeau and Eyre-Walker, 2004). These packages implement a range of tests for recombination, including: RDP, GeneConv, BootScan, MaxChi, Chimaera, SiScan, LDr^2 and LDD'. The default program values were used in all cases. Analyses were performed separately for each gene as well as for all genes concatenated together. When there were identical sequences in any analysis, only one of the sequences was included.

The substitution rates for the loci were calculated with respect to the time of divergence of *T. gondii* from *Neospora caninum*. This divergence was dated using aligned 18S rRNA sequences for the Toxoplasmatinae (Morrison *et al*., 2004). Since a likelihood-ratio test (see above) indicated that homogeneity of the substitution rates was likely ($p = 0.445$), the patristic distance to the MRCA of these two taxa was calculated via a clock-constrained maximum-likelihood tree, using the GTR+invariant+gamma model of nucleotide substitution (all parameters estimated in the model). The time of this MRCA was then estimated using a substitution rate for the 18S of 1% per 100 million years — this value was chosen because the average rate for 18S rRNA is usually considered to be 2% per 100 million years while the tree of Morrison *et al*. (2004) shows that the Toxoplasmatinae 18S have about half the average rate for the Coccidia as a whole. The TMRCA of *T. gondii* and *N. caninum* was thus estimated to be 11.3 million years ago.

Results and discussion

Earlier estimates of TMRCA of T. gondii

Estimation of the TMRCA for a set of DNA or amino acid sequences is based on the idea that the number of genetic differences (i.e. substitutions) between two sequences is directly related to the product of time and the rate at which substitutions (or mutations) have occurred. That is, if we can estimate both the number of substitutions and the substitution rate then we can calculate the amount of time that has passed since the MRCA of any two sequences (time ~ no. substitutions / substitution rate). Time can be measured in a wide range of (sometimes confusing) units, including calendar units (e.g. years, months), number of generations, number of mutations or effective population size (Ne). It is usually expressed initially in whatever units have been used to quantify the substitutions (e.g. number of nucleotide

substitutions per locus per generation), and it is then converted to the desired units using the substitution rate.

Su *et al*. (2003) used the most straightforward implementation of this method (Arbogast *et al*., 2002) by simply counting the number of observed nucleotide substitutions between the sequences. For example, to estimate the TMRCA of the atypical lineages they (correctly) counted the number of nucleotide differences between the atypical lineages and the clonal lineages, yielding an estimate of the number of substitutions that have occurred since their MRCA. That is, in all three models (Fig. 1) the MRCA of the atypical and clonal lineages is the MRCA of all of the lineages, and so this should provide an estimate of the TMRCA for all of the extant lineages.

However, in order to estimate the TMRCA of the clonal lineages they counted the number of nucleotide differences within each of the clonal lineages, which is incorrect. This calculation should be performed by counting the number of nucleotide differences *between* the clonal lineages rather than *within* them. Thus, Su *et al*. (2003) estimated the TMRCA of each lineage (which is likely to be an estimate of the time of the recent population expansions of these lineages) instead of the TMRCA of all of the clonal lineages. So, their arithmetical calculations are correct but the result has been attributed to the wrong evolutionary event.

Based on the information provided by Su *et al*. (2003) in their Supporting Online Material, the age of the MRCA of the clonal lineages can be calculated to be ≈68% of the age of the MRCA of the clonal and atypical lineages. That is, if the TMRCA for the clonal and atypical lineages is the 10^6 years suggested by Su *et al*. (2003) then the TMRCA for the clonal lineages is 0.7×10^6 years, rather than the10^4 years that they suggest. This answer supports Model (a) rather than Model (c), as favoured by Su *et al*. (2003).

Methodological issues for estimating TMRCA

Reconstruction of evolutionary events based on contemporary data is probably the most difficult form of analysis in biology. This is because there is no possibility of an independent assessment of how correct the results might be (the events are unique historical occurrences that cannot be independently observed at will, as is usually true of laboratory experimentation) and there is no "gold standard" against which to assess the relative merits of different estimation procedures. Our only recourse, then, is to assess the robustness of the results to different estimation procedures. This will at least tell us whether the data at hand appear to be sufficient for testing our hypothesis (i.e. the data are robust), and thus whether we have what is likely to be the best available estimate given the limitations of our current techniques.

With this in mind, it is worthwhile to test the three evolutionary models in more detail, in the light of the expected limitations of the calculation procedures. So, I expanded the multilocus data set of Su *et al*. (2003), and applied three different calculation procedures to these data. I thus address the question of the relative ages of the MRCA of the clonal and atypical lineages, to test whether the MRCA of the former is indeed two-thirds as old as the MRCA of the latter, as suggested by the above analysis. I then tentatively provide some actual times for the various relevant events.

There are a number of assumptions to the above analysis that may be violated for the *T. gondii* sequence data, and which it is therefore relevant to appraise. First, the calculations assume that all of the relevant mutations can be detected as observed substitutions among the contemporary sequences — in population genetics this is called the infinite sites model, which specifies that no unobserved substitutions have occurred and that all substitutions are compatible with a single evolutionary tree. This assumption is unrealistic in general, and appears to be violated by the *T. gondii* sequences. Therefore, an analysis based on a finite

sites model would be more appropriate, as these models make allowances for multiple substitutions. I have therefore used the HKY model of nucleotide substitutions in all three analyses, which seems to be the simplest model that could realistically be applied individually to each of the gene loci. When possible, I set the proportion of invariant sites to 0.80, to allow for the fact that most of the nucleotide sites are likely to be constrained to be invariant, particularly for the coding sequences.

Second, the calculations assume that the same substitution rate applies across all of the branches in the evolutionary history and across all of the genes. The latter is an unlikely assumption for most genes, and it may be particularly problematic here because some of the sequenced loci are non-coding while others are coding sequences. Variation in substitution rates between coding and non-coding sequences is a truism of sequence analysis, and this variation needs to accommodated by any analysis procedure. Furthermore, the population coalescence-time analysis (see below) makes it clear that, if nothing else, the UPRT locus has a substitution rate that is 2.5 times that of the next greatest intron-coding locus. So, in all three analyses I analysed each locus separately, and then combined the results to produce a weighted average of these individual estimates. Furthermore, homogeneity of substitution rates across evolutionary branches within the history of a gene may also be an unlikely expectation, and it can be formally tested. I used likelihood-ratio tests based on maximum-likelihood trees to perform the tests for each locus. These tests indicated that there is little evidence of unequal rates for these data (at the $p = 0.01$ level; see Nei and Kumar, 2000), as the probabilities were all in the range $p = 0.018–1.000$ (18S = 0.014, SAG1 = 0.018, NTS2 = 0.037, others = 0.158–1.000). So, this particular part of the assumption appears to be reasonably well-justified for the *T. gondii* sequence data, especially if the ribosomal sequences are excluded.

Third, the calculations assume that there has been no recombination during the evolutionary history of the organisms. This assumption is expected a priori to be violated for

the *T. gondii* data, as recombination among lineages is considered to have been a commonly occurring event in this species (Lehmann *et al*., 2004). Formally testing for recombination in sequence data can be a tricky business, as there are many genetic signatures that could be left by recombination events, and each proposed method only tests for a small subset of these (Posada and Crandall, 2001; Posada, 2002). So, it is best to use a variety of methods with different strengths and weaknesses. I used eight methods, which are in general agreement that there is little evidence of recombination within any of the loci (with the possible exception of the GRA6 locus) but that there is considerable evidence of recombination between loci (particularly for the atypical strains, notably the CASTELLS strain). Therefore, each locus must be analysed separately, and the individual results then combined. If the data were to be analysed simultaneously, and the atypical strains have been involved in recombination with the clonal lineages, then this would increase the apparent similarity of the atypical strains to the clonal strains and thus decrease the estimated difference between their TMRCAs, as well as causing apparent non-homogeneity in the estimates of substitution rates and Ne and thus violating other assumptions of the analysis.

Fourth, the calculations assume that the population size of the lineages has been constant throughout the time being investigated. This assumption is expected a priori to be violated for the *T. gondii* data, as the clonal lineages are considered to have expanded their population size dramatically over recent millennia. This issue can be formally tested, but it is not easy in this case because the DNA sequences available are not necessarily a random sample of the genetic variation within *T. gondii*, which is a necessary pre-requisite of the tests. Furthermore, the population expansion that is tested is a demographic expansion in the effective genetic population size (Ne), which refers to the genetic variation contained in an equivalent population size of randomly mating individuals. Ne is not necessarily expected to change through time if the organisms are apparently clonal, as they are for the three main *T. gondii* lineages. Thus, for *T. gondii* the population expansion is likely to be more of an

epidemiological phenomenon than a genetic phenomenon, and it is not clear how this will affect estimates of TMRCAs. So, I used one calculation method that allows the ancestral population to be of a different size to the current population sizes, for comparison.

Finally, the calculations assume that the loci have not been subjected to selection of any sort. This assumption may be true for non-coding sequences, but it is likely to be violated for many coding sequences, particularly for the surface-antigen sequences available for *T. gondii*, as these are involved in the host-parasite immune interactions. So, it will be important to consider the various types of loci separately, as they may give varying results in response to varying degrees of selection (Tanabe *et al*., 2004).

Estimating TMRCA of the clonal lineages

There are currently quite a number of available methods for estimating TMRCAs (Arbogast *et al*., 2002), but none of them implement techniques that simultaneously deal with all five of the constraints listed above. I have therefore chosen a sample of three quite different methods from what seems to be the best options: (i) a somewhat more sophisticated version of the strategy used by Su *et al*. (2003); (ii) a population-genetic method; and (iii) a phylogenetic method.

Method (i) involves calculating genetic distances between pairs of sequences for each locus separately based on a finite sites model (Arbogast *et al*., 2002). The comparisons between the clonal and atypical lineages are then based on a weighted average of these locus-specific estimates. This method differs from that of Su *et al*. (2003) in the use of a finite rather than infinite sites model to estimate the number of substitutions, and in considering each locus individually rather than first pooling the data and performing a single calculation.

It is instructive to consider the results separately for the ribosomal, housekeeping and antigen-coding loci (Table 2). With two exceptions, the ribosomal sequences gathered to date

are almost invariant among strains. This is quite unusual, especially as loci such as ITS1 are quite variable in closely related genera such as *Hammondia* and *Neospora* (Gondim *et al*., 2004), and it presumably begs for an explanation (Homan *et al*., 1997; Fazaeli *et al*., 2000b). Most of the polymorphic sites in the NTS2 locus are restricted to two regions of <200 bp each (out of 1670 bp), suggesting that there may be some common explanation for these polymorphisms. The polymorphic sites in the 18S locus mostly involve variation within strains. However, the general lack of variability means that the ribosomal sequences are not useful for estimating TMRCAs, as most of them indicate that the lineages have not yet diverged from a common ancestor.

The other two types of loci are much more variable, and therefore useful, but they do not agree on the relative ages of the MRCA of the clonal and atypical lineages. The intron loci sampled from the housekeeping genes provide an estimate that is in good agreement with that provided above (i.e. the MRCA of the clonal lineages is about two-thirds as old as the MRCA of the atypical lineages), but the (mostly) coding loci sampled from the antigen-coding genes do not detect any difference at all in the two TMRCAs. This may be a product of different selection pressures on these two types of loci, and if this is so then the results for the intron-coding sequences might be preferred (as selection pressure is likely to be less for these sequences). Note that the estimate from the data of Su *et al*. (2003) was based on the loci from both the housekeeping and antigen-coding genes, and so it should be compared only to the "Overall" estimate in Table 2, thus making it clear that the time estimates using method (i) are much greater than the previous estimate. Method (i) therefore indicates that the two-thirds estimate is likely to be an underestimate.

Method (ii) involves using the coalescent theory developed in population genetics (Hey and Nielsen, 2004). A bayesian markov chain monte carlo procedure is used to estimate demographic parameters (e.g. divergence time, effective population sizes, migration rates) under a specified model for divergence of a pair of populations from their common ancestral

population. It is not clear exactly how applicable this method is to the *T. gondii* sequences, as these data have not been randomly sampled from pre-specified populations but are instead mostly a convenience sample based on disease detection (i.e. available clinical cases). Nevertheless, the DNA sequences do represent samples of the genetic variability of the *T. gondii* lineages, and the four defined classes can be treated as four populations diverging from an ancestral population. The method thus seems to be relevant, and can be applied pairwise to the four classes. I restricted myself to the isolation model, thus excluding the possibility of migration between the populations because this would have little biological meaning under the current circumstances.

This method produces estimates for each gene of the time of coalescence of the various pairs of lineages, as well as a combined estimate for all genes, which can then be averaged (Table 3). Compared to analysis (i), this method finds less difference in estimates between the loci sampled from the housekeeping and antigen-coding genes. However, the combined estimates of the TMRCAs are close to the average estimates from the analysis using method (i). The similarities of the estimates are presumably a reflection of the robustness of the data, while the differences reflect the different assumptions of the methods. One major advantage of method (ii) over method (i) is that method (ii) allows for differences in Ne between the three populations being analysed (i.e. ancestral population and two daughter populations). If these differences exist and are correlated with the different locus types, then the TMRCA estimates from method (i) would be confounded by this variability, and thus be biased. The greater consistency of the method (ii) estimates may therefore reflect a suitable correction for changes in effective population sizes.

Method (iii) is based on explicit construction of phylogenetic trees for each locus (Arbogast *et al.*, 2002). This has similarities to method (i) in that it involves calculating genetic divergences between pairs of sequences based on a finite sites model, but it differs in that it takes into account the non-independence of the divergences due to common branches

connecting the sequences on the phylogenetic tree. The TMRCA is thus based directly on the branch lengths on the tree, called patristic distances. The comparisons between the clonal and atypical lineages can then be based on a weighted average of these locus-specific estimates.

The TMRCA estimates using this method are all much greater than for the other two methods (Table 4), but otherwise the results are similar to those from method (i). As found using method (i), there is a detectable difference between the estimates based on the loci sampled from the housekeeping and antigen-coding genes, with the latter finding little difference in TMRCA for the clonal and atypical lineages. I also tried an alternative strategy, in which the data for the different loci were first combined and then a single phylogenetic tree produced, from which the patristic distances were estimated (a "pooled-gene analysis"). This approach suffered from a severe loss of sequence information, as the tree was reduced to only five taxa (the three clonal lineages and two atypical lineages). Nevertheless, it produced results with reduced TMRCAs, which were thus more in accord with those from methods (i) and (ii).

Overall, I conclude that the estimated TMRCAs are reasonably robust to the form of data analysis. Using methods (i) and (iii), the calculations seem to be affected by differential selection pressure among loci and by changing Ne through time, but these effects are not large enough to alter the overall consistency of the TMRCA estimates. In any event, the data and analyses allow us to distinguish clearly between the three evolutionary models. The estimated TMRCA of the three clonal lineages is probably ~80% as old as the TMRCA of the atypical lineages, and possibly more. This is inconsistent with Model (c). Furthermore, it would require quite dramatic substitution rates in all of the lineages to make the data compatible with Model (b). Therefore, Model (a) is the one that is most consistent with the data analysed here.

Age of the lineages

This brings us to the topic of exactly what age the MRCA of all of the *Toxoplasma* lineages might be. Su *et al*. (2003) provide a range of estimates, based on sequence substitution rates for *Plasmodium falciparum* as well as their own estimate for *T. gondii*. Their preferred estimate of the age of *T. gondii* is biased towards use of the *P. falciparum* rates; however, it is not clear that an intra-species rate can be meaningfully based on the rate calculated for a taxon that shares a common ancestor at least several hundred millions of years ago. So, the most straightforward procedure is to use solely the data that are currently available for *T. gondii*, with whatever limitations they might entail.

Su *et al*. (2003) provide data for estimating (based on the time of the divergence of *T. gondii* and *N. caninum*) the substitution rate for the introns of three of the *T. gondii* housekeeping genes (ACT1, ATUB, MIC2), and it is currently possible to perform comparable calculations for one non-coding ribosomal locus (NTS2; data from Fazaeli *et al*., 2000b) and the coding region of an antigen gene (SAG1; based on multiple database sequences). The estimate of the SAG1 substitution rate is slightly less than those for the three introns, while the NTS2 rate is only one-third of these rates. Given this variation, it is probably best to restrict the age estimate to the housekeeping introns, where the rate estimates are replicated across loci. Taking the geometric mean of these rates (2.4×10^{-8}) and applying it to my data for the housekeeping genes, the estimated ages for the MRCA of the lineages studied here are 128,000 years based on the distance estimates (Table 2), 123,000 years for the coalescence times (Table 3) and 163,000 years for the patristic distances (Table 4).

This approach provides a minimum estimate, of course, based on the lineages that have been sequenced so far — other extant lineages may provide an older age estimate. Furthermore, Grigg *et al*. (2001) and Grigg and Suzuki (2003) suggest that the extant lineages are all derived from two ancestral gene pools (which they refer to as ‘Adam’ and ‘Eve’). This

implies that there has been a severe population bottleneck at some time in the history of *T. gondii*, through which all of the current lineages have passed (that is, the lineages all coalesce at that time), or possibly that the global population was chronically small before that time. If this scenario is correct, then clearly the date estimated here refers to the time of the bottleneck, rather than to the time of origin of the two ancestral gene pools. The origin of the species itself, involving its differentiation from closely related species such as *Hammondia hammondi* and *Neospora caninum*, is presumably much further back in the past (estimated here as 11 million years ago), and the evolutionary history between that time and the bottleneck probably cannot be deduced using genetic data from extant organisms.

Interestingly, the evolutionary model that this scenario suggests for *T. gondii* (i.e. apparent clonality due to recent expansion of a few relatively ancient lineages) is rather similar, in both structure and timing, to that proposed by Joy *et al.* (2003) for *P. falciparum* (see also Tanabe *et al.*, 2004). Indeed, the timing of the MRCA of both species coincides with the third "out of Africa" population expansion of *Homo sapiens*, which is dated as 80,000–150,000 years ago (Templeton, 2002). This coincidence may favour a model in which these parasites existed at a chronically low population level for the vast majority of their evolutionary history, only coming to global prominence as a result of their association with humans.

It is also possible to derive an estimate for the time of population expansion of the clonal lineages, by adapting the techniques used above for timing the origin of the clonal lineages. This approach relies exclusively on the data for clonal lineage I, as it is the only one with a sufficient number of variable sequences in the data set used here — at the best of times, the dating of relatively recent evolutionary events is unreliable using phylogenetic techniques, because the amount of sequence polymorphism is so small, and it cannot be done at all if there is no genetic variation. The estimated time of expansion is 6,000 years based on the genetic distance estimates, 8,000 years for the coalescence times and 11,000 years for the

patristic distances. These estimates accord reasonably well with the presumed rapid demographic expansion of the human population over the past 10,000 years (Cavalli-Sforza *et al.*, 1994), with the associated extensive domestication of many of the intermediate hosts of *T. gondii*. It thus seems likely that the expansion of the clonal lineages is a direct consequence of the demographic expansion of their human-associated hosts. This phenomenon has been noted for many other parasitic species (Morrison and Höglund, 2005), as well.

The main limitation of this study is the apparently biased nature of the sequence sampling. Accurate reconstruction of evolutionary events requires that all of the relevant lineages be sampled. The key lineages in this instance are those that branch near the base of the various study groups, as these are the ones that most influence the reconstruction of the oldest ancestors; so, the accurate estimation of the TMRCA for a group of lineages requires that data be collected from as many basal lineages as possible. Most of the currently available *T. gondii* strains are derived from human clinical cases, with most of the rest being from diseased domesticated animals. It is unlikely that these strains provide a sufficient representation of the basal lineages, and thus the oldest MRCA time quoted here is probably an under-estimate. A more accurate estimate of the true time will only become available when data are collected from non-domesticated animals and possibly also asymptomatic individuals (cf. Ajzenberg *et al.*, 2004), as it is more likely that these will characterize the basal lineages. Such data are becoming available for microsatellites, and it may be worthwhile to pursue similar analysis of these data, in addition to the sequence data that I have used here. In the absence of fossil evidence for this or any other apicomplexan species, such estimates will have to suffice.

Acknowledgements

Thanks to Daniel Ajzenberg for first suggesting this topic to me, and to Daniel Ajzenberg and Chunlei Su for helpful discussions.

References

Ajzenberg D, Bañuls A-L, Tibayrenc M, Dardé ML (2002). Microsatellite analysis of *Toxoplasma gondii* shows considerable polymorphism structured into two main clonal groups. Int J Parasitol 32: 27–38.

Ajzenberg D, Bañuls AL, Su C, Dumètre A, Demar M, Carme B, Dardé ML (2004). Genetic diversity, clonality and sexuality in *Toxoplasma gondii*. Int J Parasitol 34: 1185–1196.

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002). Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu Rev Ecol Syst 33: 707–740.

Barry AE, Leliwa A, Choi M, Nielsen KM, Hartl DL, Day KP (2003). DNA sequence artifacts and the estimation of time to the most recent common ancestor (TMRCA) of *Plasmodium falciparum*. Mol Biochem Parasitol 130: 143–147.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). The History and Geography of Human Genes. Princeton Uni Press, Princeton, NJ.

Edwards SV, Beerli P (2000). Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54: 1839–1854.

Fazaeli A, Carter PE, Dardé ML, Pennington TH (2000a). Molecular typing of *Toxoplasma gondii* strains by *GRA6* gene sequence analysis. Int J Parasitol 30: 637–642.

Fazaeli A, Carter PE, Pennington TH (2000b). Intergenic spacer (IGS) polymorphism: a new genetic marker for differentiation of *Toxoplasma gondii* strains and *Neospora caninum*.

J Parasitol 86: 716–723.

Gondim LFP, Laski P, Gao L, McAllister MM (2004). Variation of the internal transcribed spacer 1 sequence within individual strains and among different strains of *Neospora caninum*. J Parasitol 90: 119–122.

Grigg ME, Bonnefoy S, Hehl AB, Suzuki Y, Boothroyd JC (2001). Success and virulence in *Toxoplasma* as the result of sexual recombination between two distinct ancestries. Science 294: 161–165.

Grigg ME, Suzuki Y (2003). Sexual recombination and clonal evolution of virulence in *Toxoplasma*. Microbes Infect 5: 685–690.

Hey J, Nielsen R (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.

Homan WL, Limper L, Verlaan M, Borst A, Vercammen M, van Knapen F (1997). Comparison of the internal transcribed spacer, ITS 1, from *Toxoplasma gondii* isolates and *Neospora caninum*. Parasitol Res 83: 285–289.

Howe DK, Sibley LD (1994). *Toxoplasma gondii*: analysis of different laboratory stocks of the RH strain reveals genetic heterogeneity. Exp Parasitol 78: 242–245.

Johnson AM (1997). Speculation on possible life cycles for the clonal lineages in the genus *Toxoplasma*. Parasitol Today 13: 393–397.

Joy DA, Feng X, Mu J, Furuya T, Chotivanich K, Krettli AU, Ho M, Wang A, White NJ, Suh E, Beerli P, Su X-Z (2003). Early origin and recent expansion of *Plasmodium falciparum*. Science 300: 318–321.

Lehmann T, Blackston CR, Parmley SF, Remington JS, Dubey JP (2000). Strain typing of *Toxoplasma gondii*: comparison of antigen-coding and housekeeping genes. J Parasitol 86: 960–971.

Lehmann T, Graham DH, Dahl ER, Bahia-Oliveira LMG, Gennari SM, Dubey JP (2004).

Variation in the structure of *Toxoplasma gondii* and the roles of selfing, drift, and epistatic selection in maintaining linkage disequilibria. Infect Genet Evol 4: 107–114.

Luton K, Gleeson M, Johnson AM (1995). rRNA gene sequence heterogeneity among *Toxoplasma gondii* strains. Parasitol Res 81: 310–315.

Martin D, Williamson C, Posada S (2005). RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21: 260–262.

Morrison DA, Bornstein S, Thebo P, Wernery U, Kinne J, Mattsson JG (2004). The current status of the small subunit rRNA phylogeny of the coccidia (Sporozoa). Int J Parasitol 34: 501–514.

Morrison DA, Höglund J (2005). Testing the hypothesis of recent population expansions in nematode parasites of human-associated hosts. Heredity (in press).

Nei M, Kumar S (2000). Molecular Evolution and Phylogenetics. Oxford Uni. Press, Oxford.

Piganeau G, Eyre-Walker A (2004). A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. Heredity 92: 282–288.

Posada D (2002). Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol Biol Evol 19: 708–717.

Posada D (2003). Selecting models of evolution. In: The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny (Salemi M, Vandamme A-M, eds). Cambridge Uni Press, Cambridge, pp 256–282.

Posada D, Crandall KA (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA 98: 13757–13762.

Sibley LD (2003). Recent origins among ancient parasites. Vet Parasitol 115: 185–198.

Su C, Evans D, Cole RH, Kissinger JC, Ajioka JW, Sibley LD (2003). Recent expansion of *Toxoplasma* through enhanced oral transmission. Science 299: 414–416.

Swofford DL (2002). PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Ver. 4. Sinauer Associates, Sunderland, MA.

Tanabe K, Sakihama N, Hattori T, Ranford-Cartwright L, Goldman I, Escalante AA, Lal AA (2004). Genetic distance in housekeeping genes between *Plasmodium falciparum* and *Plasmodium reichenowi* and within *P. falciparum*. J Mol Evol 59: 687–694.

Templeton AR (2002). Out of Africa again and again. Nature 416: 45–51.

Table 1. Nucleotide sequences used to estimate the relative ages of the *Toxoplasma* lineages.

Abbreviation	Locus description (sequence type)	No. sequences (clonal/atypical)	Average length (bases)	Reference
Ribosomal genes				
18S	small subunit ribosomal RNA	10/1	1747.5	Luton *et al.* (1995)
ITS1	internal transcribed spacer 1	18/12	393	Su *et al.* (2003)
5.8S	5.8S ribosomal RNA	3/2	159	
ITS2	internal transcribed spacer 2	3/2	341	
5S	5S ribosomal RNA	17/6	109.5	Fazaeli *et al.* (2000b)
NTS2	non-transcribed spacer 2	16/5	1669.5	Fazaeli *et al.* (2000b)
Housekeeping genes				
ACT1	actin subunit 1 (intron)	22/7	445.5	Lehmann *et al.* (2000); Su *et al.* (2003)
ATUB	α-tubulin (intron)	22/7	585.5	Lehmann *et al.* (2000); Su *et al.* (2003)
BTUB	β-tubulin (intron)	22/7	614.5	Lehmann *et al.* (2000); Su *et al.* (2003)
FOL1	dihydrofolate reductase-thymidylate synthase (intron)	13/4	562	Lehmann *et al.* (2000)
MIC2	micronemal protein 2 (intron)	11/3	955	Su *et al.* (2003)
TgMA	myosin subunit A (intron)	25/14	160	Ajzenberg *et al.* (2004)
UPRT	uracil phosphoribosyl transferase (intron)	11/3	926	Su *et al.* (2003)
Antigen-coding genes				
B10	excreted-secreted protein of unknown function (intron)	13/4	572	Lehmann *et al.* (2000)
BSR4	bradyzoite surface antigen 4 (coding)	24/5	1173	Grigg *et al.* (2001); Grigg and Suzuki (2003)
GRA6	dense granule antigen 6 (coding)	27/5	740	Fazaeli *et al.* (2000a)
MAG1	bradyzoite surface antigen 1 (intron + coding)	13/4	1279	Lehmann *et al.* (2000)
SAG1	tachyzoite surface antigen 1 (coding)	7/2	801	
SAG2	tachyzoite surface antigen 2 (coding)	16/7	1281.5	Lehmann *et al.* (2000)
SAG4	tachyzoite surface antigen 4 (coding)	10/5	519	Grigg *et al.* (2001); Grigg and Suzuki (2003)

Table 2. Average pairwise genetic distance (substitutions per site) between *Toxoplasma* gene sequences for different groupings of the sequences, based on the HKY+invariant model of sequence evolution.

Locus	Average genetic distance between sequences			Ratio of 'among clonal' to 'between clonal and atypicals'
	Within clonal lineages	Among clonal lineages	Between clonal lineages and atypicals	
Ribosomal genes				
18S	0.00138	0.00223	0.00476	0.47
ITS1	0.00000	0.00000	0.00087	
5.8S	0.00000	0.00000	0.00000	
ITS2	0.00000	0.00000	0.00000	
5S	0.00000	0.00000	0.00000	
NTS2	0.00092	0.00249	0.00297	0.84
Housekeeping genes				
ACT1	0.00000	0.00579	0.00715	0.81
ATUB	0.00000	0.00137	0.00178	0.77
BTUB	0.00000	0.00579	0.00551	1.05
FOL1	0.00000	0.00406	0.00459	0.89
MIC2	0.00013	0.00168	0.00589	0.28
TgMA	0.00061	0.00561	0.01134	0.50
UPRT	0.00082	0.00921	0.01352	0.68
Antigen-coding genes				
B10	0.00000	0.00722	0.00856	0.84
BSR4	0.00006	0.02890	0.02908	0.99
GRA6	0.00041	0.01474	0.01306	1.13
MAG1	0.00036	0.00284	0.00220	1.29
SAG1	0.00125	0.00960	0.00996	0.96
SAG2	0.00000	0.00536	0.00511	1.05
SAG4	0.00032	0.02397	0.02199	1.09
Averages				
Overall	0.00050	0.00769	0.00861	0.89
Housekeeping	0.00023	0.00477	0.00710	0.67
Antigen	0.00032	0.01250	0.01216	1.03

Table 3. Average pairwise coalescence time to the MRCA (substitutions per locus) between *Toxoplasma* gene sequences for different groupings of the sequences, based on the HKY model of sequence evolution and the isolation demographic model.

Locus	Average time to MRCA		Ratio of 'among clonal' to 'between clonal and atypicals'
	Among clonal lineages	Between clonal lineages and atypicals	
Housekeeping genes			
ACT1	1.97	2.43	0.81
ATUB	1.77	2.50	0.71
BTUB	2.01	3.10	0.65
FOL1	1.88	2.73	0.69
MIC2	1.63	1.88	0.87
TgMA	1.59	2.55	0.62
UPRT	1.39	1.25	1.11
Antigen-coding genes			
B10	4.13	10.89	0.38
BSR4	9.22	5.03	1.83
GRA6	3.94	5.33	0.74
MAG1	2.94	6.33	0.46
SAG1	1.88	6.40	0.29
SAG2	5.97	6.93	0.86
SAG4	5.51	9.15	0.60
Combined			
Housekeeping	2.64	3.79	0.70
Antigen	5.50	6.74	0.82

Table 4. Patristic distance to the MRCA (substitutions per site) between *Toxoplasma* gene sequences for different groupings of the sequences, based on the HKY+invariant model of sequence evolution and maximum-likelihood tree-building.

Locus	Largest distance to MCRA		Ratio of 'among clonal' to 'between clonal and atypicals'
	Among clonal lineages	Between clonal lineages and atypicals	
Ribosomal genes			
18S	0.00173	0.00260	0.66
ITS1	0.00000	0.00260	
5.8S	0.00000	0.00000	
ITS2	0.00000	0.00000	
5S	0.00000	0.00000	
NTS2	0.00222	0.00222	1.00
Housekeeping genes			
ACT1	0.00421	0.00421	1.00
ATUB	0.00110	0.00110	1.00
BTUB	0.00410	0.00410	1.00
FOL1	0.00334	0.00334	1.00
MIC2	0.00082	0.00395	0.21
TgMA	0.00580	0.00783	0.74
UPRT	0.00783	0.00783	1.00
Antigen-coding genes			
B10	0.00694	0.00694	1.00
BSR4	0.01623	0.01623	1.00
GRA6	0.00991	0.00991	1.00
MAG1	0.00235	0.00235	1.00
SAG1	0.00679	0.00679	1.00
SAG2	0.00392	0.00430	0.91
SAG4	0.01486	0.01730	0.86
Averages			
Overall	0.00635	0.00683	0.93
Housekeeping	0.00374	0.00452	0.83
Antigen	0.00809	0.00837	0.97
Pooled-gene analysis			
Overall	0.00610	0.00674	0.91
Housekeeping	0.00307	0.00413	0.74
Antigen	0.00813	0.00881	0.92

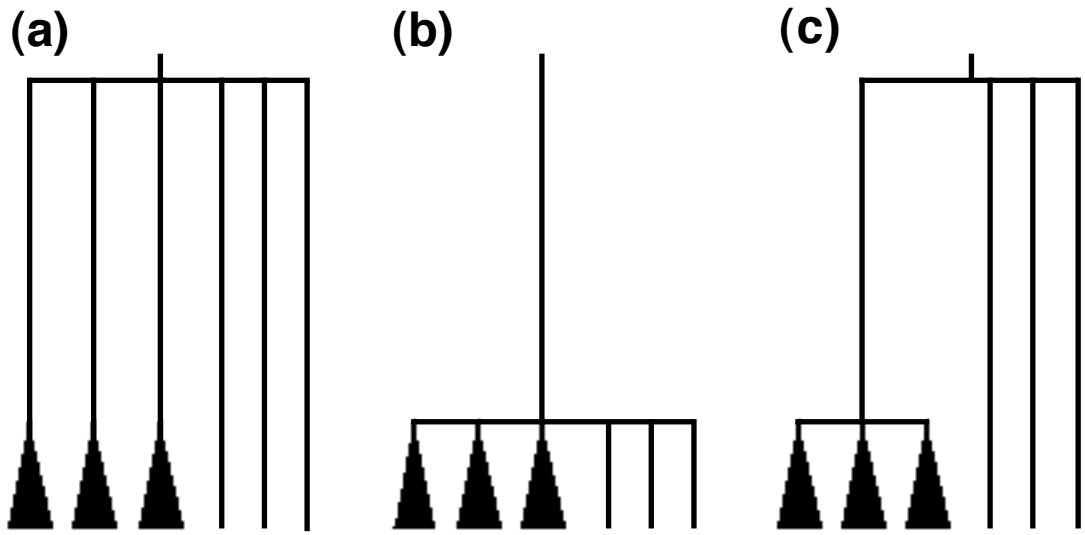

Fig. 1. Schematic representation of three proposed models for the evolutionary history of *T. gondii*, (**a**) and (**b**) from Grigg and Suzuki (2003) and (**c**) from Su *et al*. (2003). The most recent common ancestor (MCRA) is at the top of each diagram with time proceeding towards the bottom, the filled triangles represent the recent population expansions of the three clonal lineages, and the other three lineages represent examples of the "atypical" strains. In all cases, the proposed recombination between lineages (which would be represented by horizontal lines connecting the lineages) has been omitted for clarity.