Third generation sequencing revises the molecular karyotype for *Toxoplasma gondii* and identifies emerging copy number variants in sexual recombinants.

Jing Xia[1], Aarthi Venkat[2,3], Michael L. Reese[4], Karine Le Roch[5], Ferhat Ay[3,5] and Jon P. Boyle[1*]

[1]Department of Biological Sciences, Dietrich School of Arts and Sciences, University of Pittsburgh

[2]Computational Biology and Bioinformatics, Yale University, New Haven, CT, 06520.

[3]La Jolla Institute for Immunology, La Jolla, CA, 92037.

4University of Texas-Southwestern, Dallas, TX. 75390

[5]Department of Molecular, Cell and Systems Biology, College of Agricultural and Life Sciences, University of California-Riverside.

[5]Department of Pediatrics, University of California - San Diego, La Jolla, CA, 92093.

*Author to whom correspondence should be addressed:

Dietrich School of Arts and Sciences

Department of Biological Sciences

University of Pittsburgh

Life Sciences Annex 101

Pittsburgh, PA. 16260

email: boylej@pitt.edu

ABSTRACT

Toxoplasma gondii is an obligate intracellular parasite that has a significant impact on human health, especially in the immunocompromised. This parasite is also a useful genetic model for intracellular parasitism given its ease of culture in the laboratory and relevant animal models. However, as for many other eukaryotes, the *T. gondii* genome is incomplete, containing hundreds of sequence gaps due to the presence of repetitive and/or uncloneable sequences that prevent complete telomere-to-telomere de novo chromosome assembly. Here, we report the first use of single molecule DNA sequencing to generate near complete de novo genome assemblies for *T. gondii* and its near relative, *N. caninum*. Using the Oxford Nanopore Minion platform, we dramatically improved the contiguity of the *T. gondii* genome (N50 of ~6.6Mb) and increased overall assembled sequence compared to current reference sequences by ~2 Mb. Multiple complete chromosomes were fully assembled as evidenced by clear telomeric repeats on the end of each contig. Interestingly, for all of the *Toxoplasma gondii* strains that we sequenced (RH, CTG, II×III F1 progeny clones CL13, S27, S21, and S26), the largest contig ranged in size between 11.9 and 12.1 Mb in size, which is larger than any previously reported *T. gondii* chromosome. This was due to a repeatable and consistent fusion of chromosomes VIIb and VIII. These data were further validated by mapping existing *T. gondii* ME49 Hi-C data to our assembly, providing parallel lines of evidence that the *T. gondii* karyotype consists of 13, rather than 14, chromosomes. In addition revising the molecular karyotype we were also able to resolve hundreds of repeats derived from both coding and non-coding tandem sequence expansions. For well-known host-targeting effector loci like rhoptry protein 5 (ROP5) and ROP38, we were also able to accurately determine the precise gene count, order and orientation using established assembly approaches and the most likely primary sequence of each using our own assembly correction scripts tailored to correcting homopolymeric run errors in tandem sequence arrays. Finally, when we compared the *T. gondii* and *N. caninum* assemblies we found that while the 13 chromosome karyotype was conserved, we determined

53 that previously unidentified large scale translocation events occurred in *T. gondii* and *N.*

54 *caninum* since their most recent common ancestry.

55

56 **INTRODUCTION**

57 *Toxoplasma gondii* and its Apicomplexan relatives are highly successful animal pathogens,
58 infecting a wide variety of warm-blooded animals including humans and domesticated animals.
59 *T. gondii* infection can lead to severe toxoplasmosis in immunocompromised individuals and in
60 congenitally-infected fetuses (Joynson and Wreghitt 2005), and is a leading cause of blindness
61 due to its ability to infect the eye causing ocular toxoplasmosis (Jones and Holland 2010). *T.*
62 *gondii* belongs to the phylum Apicomplexa, a large group of animal and human pathogens
63 including *Neospora*, *Eimeria*, *Plasmodium* and *Cryptosporidium*. The ease of genetic
64 manipulation, accessibility to cellular and biochemical experiments, and well-established animal
65 model make *T. gondii* an important system for studying the biology of Apicomplexans (Kim and
66 Weiss 2004). Genomic analysis tools for this organism have been under development for
67 decades. Data housed at ToxoDB.org, the primary genomic repository for *T. gondii* genome-
68 wide data, presently includes sequence, *de novo* assemblies and annotation of multiple *T.*
69 *gondii* genomes, next-generation sequence data for an additional 60 *T. gondii* genomes, as well
70 as draft assemblies for both *Hammondia hammondi* and *Neospora caninum* (Lorenzi *et al.*
71 2016a).

72 Availability of a complete reference genome that contains accurate representations of all
73 small- or large-scale structural variants is essential to have a better understanding of gene
74 content, genotype-phenotype relationships, and the evolution of unique traits in parasites of
75 humans and other animals. However, like all eukaryotic genomes, a substantial part of the *T.*
76 *gondii* genome consists of repetitive elements (Matrajt *et al.* 1999), making gap-free *de novo*
77 assembly impossible using standard 1st or 2nd generation sequencing approaches. Even with
78 exceptionally high coverage, these approaches fail to resolve repetitive regions or complex
79 structural variants with repeat units that are larger than the size of the individual reads. Three of
80 the *T. gondii* reference genomes in ToxoDB (Gajria *et al.* 2008) were constructed by combining
81 high quality 1st generation Sanger (Sanger *et al.* 1977) and 2nd generation 454 (Roche Applied

82 Science) sequence data, yet these genomes still have hundreds of sequence gaps of unknown
83 sequence content and length. Assembly gaps typically mask repetitive regions which can
84 contain previously unknown protein-coding genes, additional copies of genes found in tandem
85 gene arrays, and in some cases, they may also lead to incorrect predictions of chromosomal
86 structure. This problem is not unique to *T. gondii* and other apicomplexan genomes. For
87 example, all versions of the human genome have thousands of gaps due to incorrect assembly
88 of repetitive sequence (Vollger *et al.* 2019), including assemblies generated recently using new
89 sequencing technologies like those applied here.

90 In recent years single molecule sequencing approaches (developed by Oxford Nanopore
91 and PacBio) have revolutionized *de novo* sequence assembly by enabling high-throughput
92 generation of kilobase-sized sequence reads. These approaches have allowed for resolution of
93 the vast majority of repeat-driven sequence assembly gaps, the detection and assembly of
94 previously intractable structural variants with species, and when combined with 2nd generation
95 sequencing data can be used to generate near-complete *de novo* genome assemblies with high
96 (>99%) nucleotide accuracy. Indeed, whole genome assemblies of several organisms including
97 bacteria (Madoui *et al.* 2015; Fournier *et al.* 2017; Diaz-Viraque *et al.* 2018), parasites (Lapp *et*
98 *al.* 2018), plants (Schmidt *et al.* 2017; Michael *et al.* 2018), and mammals (Jain *et al.* 2018)
99 using a such a hybrid approach have been reported, generating assemblies of unprecedented
100 contiguity. Here, we apply Oxford Nanopore sequencing and de novo assembly using the
101 Minion Platform to multiple isolates of *T. gondii*, F1 progeny of a cross between two canonical *T.*
102 *gondii* strains, and one of its nearest extant relatives, *Neospora caninum*. Using these data, we
103 have generated mostly gap-free, telomere-to-telomere assemblies for the majority of the *T.*
104 *gondii* chromosomes from each sequenced isolate and/or species. These assemblies revise the
105 molecular karyotype for *T. gondii*, formally establishing that this parasite has 13, rather than 14,
106 chromosomes. This result was confirmed by Hi-C sequencing technologies as described
107 previously (Bunnik *et al.* 2019).This karyotype is conserved across all *T. gondii* parasite strains

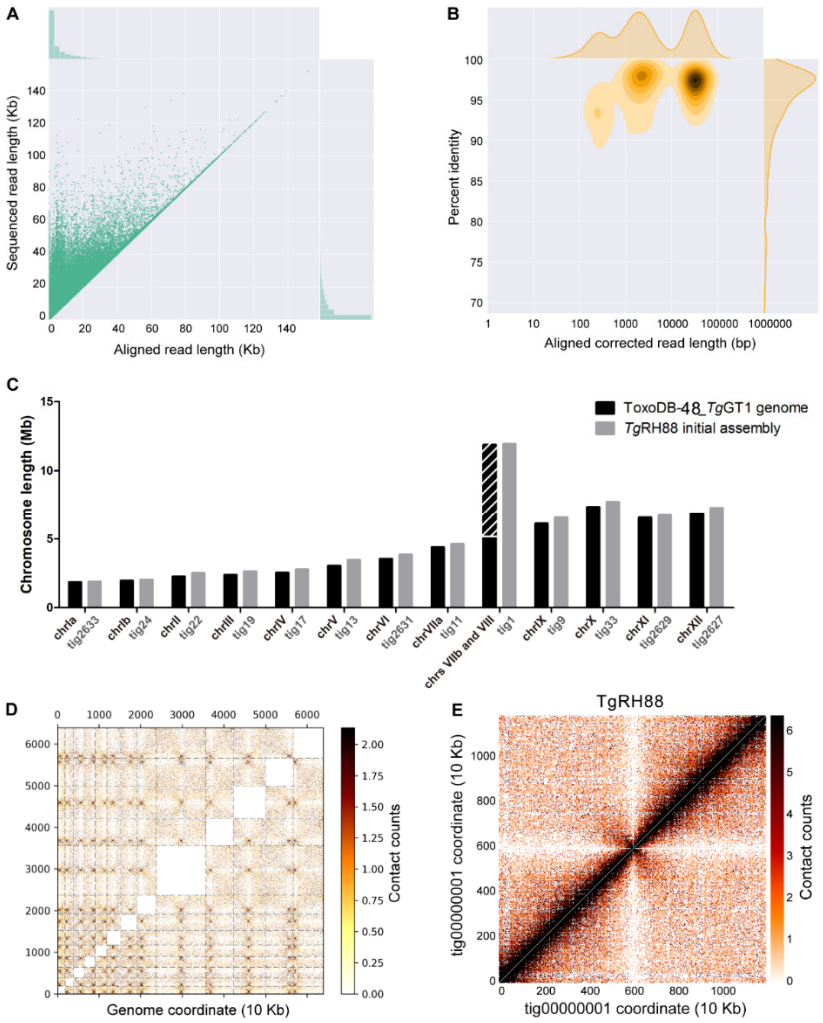
108 that we sequenced, and is also conserved in *N. caninum*. Our new assemblies have also
109 resolved copy number estimates at simple- and complex repetitive loci, including multiple
110 tandem gene arrays harboring known host-targeting effector genes. Moreover, by directly
111 assembling genomes from multiple F1 progeny experimental crosses, we have determined for
112 the first time that changes in gene copy number occur during sexual recombination, and that
113 this occurs frequently at some *T. gondii* loci (like the ROP5 locus) and infrequently at others
114 (like the MAF1 locus), and may explain some of the differences in pathogenicity between *T.*
115 *gondii* strains. Finally, our *de novo* assembly of the *N. caninum* Liverpool genome shows that
116 the current draft assembly of this organism dramatically overestimates chromosome-level
117 synteny with *T. gondii*. This discovery is also corroborated in multiple *N. caninum* strains in a
118 companion paper (See cover letter for details on co-submitted manuscript).
119

120 **RESULTS**

121 ***De novo* assembly of *Tg*RH88 genome using nanopore reads revises the *T. gondii***
122 **karyotype**

123 The majority of the *T. gondii* isolates collected from North America and Europe belong to three
124 predominant clonal lineages, types I, II, and III (Sibley and Ajioka 2008) and RH strain is a
125 representative strain of the type I lineage (Pfefferkorn and Pfefferkorn 1976). While RH strain
126 shows some unique phenotypes including higher growth rate *in vitro* and inability to form cysts
127 (Villard *et al.* 1997; Khan *et al.* 2009) and is most frequently used in *T. gondii* studies, a
128 complete *de novo* assembly for *Tg*RH genome is lacking. Therefore, our initial efforts were to
129 use long read single molecule sequencing to generate a more complete *Tg*RH88 assembly.
130 To take advantage of the long-read technology, high-molecular weight genomic DNA of
131 *Tg*RH88 strain was extracted using an optimized protocol which was originally designed for
132 extraction of gram-negative bacteria and mammalian cell DNA (Quick 2018). This protocol
133 included a phenol-chloroform-isoamyl alcohol extraction followed by ethanol precipitation, and

134 the use of large bore pipette tips for all manipulations (see Materials and Methods). For
135 TgRH88, we obtained 24 µg of genomic DNA from 2×10^8 tachyzoites, and 400 ng was used for
136 MinION sequencing. A 48-hour sequencing run on a single flow cell yielded 648,491 reads
137 containing 7.40 Gb of sequences for *Tg*RH88 genome. Assuming a 65 Mb *T. gondii* genome
138 size, these reads represented a genome coverage of ~114×. While more detailed sequence
139 metrics are described below, the sequence reads we obtained robustly aligned to the TgGT1
140 reference genome found at ToxoDB.org (**Figure 1A**).



141

142 ***Figure 1. Primary de novo assembly of TgRH88 genome using nanopore***
143 ***reads revises T. gondii karyotype.*** *(A) Bivariate plot showing a comparison*
144 *of the aligned read length with the sequencing read length. (B) Bivariate plot*
145 *showing a comparison of the aligned corrected read length (log10*
146 *transformed) with the percent identity. In this case corrected reads refer to the*

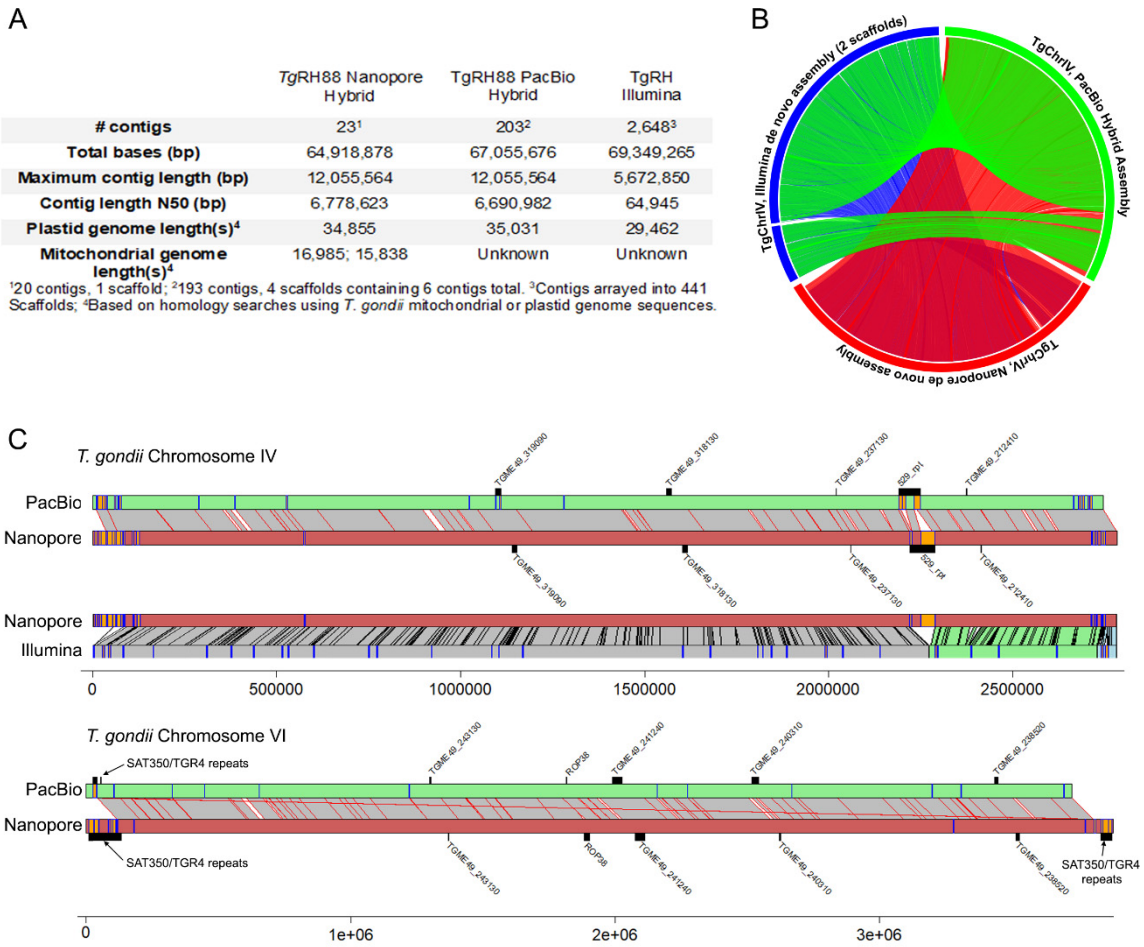
method deployed by Canu using read overlap. (C) Histogram showing comparison of chromosome size between ToxoDB-48_TgGT1 genome and TgRH88 initial long-read assembly. (D) Interchromosomal Hi-C contact-count heat map plotted using the TgRH88 initial long-read assembly sequence showing 13 chromosomes in the assembly. (E) Intrachromosomal Hi-C contact-count heat map plotted using the sequence of TgRH88_tig00000001 in TgRH88 initial long-read assembly showing no aberrant signal along the contig

Before assembly, Canu v1.7.1 corrected all reads > 1000 bp, and after alignment we found that 99.97% of the corrected *Tg*RH88 reads could be mapped to the ToxoDB-48_*Tg*GT1 reference genome, and the mean percent identity was 92.9% (**Figure 1B**). The resulting Canu-corrected reads were subjected to *de novo* assembly using Canu v1.7.1, which yielded a *Tg*RH88 primary assembly with a size of 64.40 Mb. The assembly consisted of 23 contigs and 6 of them contained telomeric repeat sequences (TTTAGGG or CCCTAAA) on at least one end. Interestingly, by aligning the *Tg*RH88 assembly sequences to the ToxoDB-48_*Tg*GT1 reference genome, we noticed that the sequences annotated as chrVIIb and chrVIII were parts of a single contig in our *Tg*RH88 assembly (TgRH88_tig00000001; **Figure 1C**). This contig, which was 11.93 Mb in length, was longer than any previously reported *T. gondii* chromosome, and suggested to us that *T. gondii* consists of 13 (instead of 14) chromosomes, with the chrVIIb and chrVIII being actually a single chromosome. Given that prior work using Hi-C chromosome conformation capture sequencing suggested a similar fusion between chromosomes VIIb and VIII (Bunnik *et al.* 2019), we mapped the Hi-C reads from that study onto our TgRH88 de novo assembly to determine if it has similar contact counts. As shown in **Figure 1D** the Hi-C data identified the position of 13, rather than 14, interchromosomal contact points (representing centromeres; (Bunnik *et al.* 2019)) and an intrachromosomal contact map across TgRH88tig00000001 indicating that this did indeed represent a single contiguous chromosome (**Figure 1E**). These findings provide direct assembly-based evidence that sequence fragments previously referred to as distinct chromosomes (VIIb and VIII) were in fact two parts of the same chromosome. We have named this contig TgRH88_tig00000001_chrVIII.

176 The RH strain is one of the most commonly used laboratory strains given its genetic
177 tractability and robust *in vitro* growth characteristics (Saeij *et al.* 2005), but at the genomic level
178 this strain has been subject to much less formal annotation compared to strain types GT-1,
179 ME49 and VEG. However, we did identify two additional RH strain de novo assemblies in
180 Genbank, one generated using Illumina technology (Lau *et al.* 2016); Genbank BioProject
181 PRJNA294483) and the other using single-molecule long read sequencing using PacBio RS
182 technology (BioProject PRJNA279557). As expected, our Nanopore and the PacBio RS
183 assembly were more contiguous than the de novo Illumina assembly, having similar contig N50
184 values (6.7 Mb for both; **Figure 2A**) and both predicting the ~12 Mb chromosome VIII. Moreover
185 total size of the predicted nuclear-encoded genomes in each assembly differ by only 0.8% (a
186 raw difference of 513,977 bp of additional sequence present in the Nanopore assembly;
187 discussed further below). Despite this congruence of the two long-read assemblies, the most
188 striking observed difference between them is the total number of contigs (23 vs. 203,
189 respectively) in each of the final assemblies (prior to scaffolding; **Figure 2A**). Based on Nucmer
190 alignments, ~362 Kb of this unplaced sequence is shared between the two assemblies, while
191 the remaining sequences (~70 and 2700 Kb for the Nanopore and PacBio RH assemblies,
192 respectively) were unshared. We also used blast to identify contigs harboring what appeared to
193 be sequence derived from either of the two organellar genomes of *T. gondii*, the apicoplast and
194 the mitochondrion and compared them to the other existing RH assemblies (Figure 2A). One
195 contig ~35 Kb in size from each assembly was derived from the plastid genome, which is the
196 expected size for this organellar genome (Lorenzi *et al.* 2016b). For the mitochondrial genome,
197 we identified two non-chromosomal contigs of ~15 Kb in size in our assembly that likely
198 harbored at least fragments of this genome (Figure 2A), while there were no contigs in the other
199 RH assemblies with any resemblance to a mitochondrial genome. It is likely that this is due to
200 pre-filtering genome assemblies prior to submission to genbank to eliminate mitochondrial
201 sequences. However for all of our Nanopore assemblies presented here and submitted to

202 genbank, we have annotated all putative mitochondrial genome contigs with the
203 "location=mitochondria" tag so that they would pass the filters in Genbank. However given the
204 complexities of the mitochondrial genome and its assembly for *T. gondii* (as evidenced by
205 recent work from the Kissinger group;
206 https://www.biorxiv.org/content/10.1101/2020.05.16.099366v1), we did not perform any
207 extensive analyses of these sequences.
208 To further compare our assembly to the two existing RH88 de novo assemblies we
209 compared them using Mummer (Delcher *et al.* 2003) and Tandem Repeats Finder (Benson
210 1999). While complete chromosomes aligned with high synteny between the Nanopore and
211 PacBio assemblies (Figure 2B), chromosomes in the Illumina only RH assembly were
212 distributed across multiple scaffolds (e.g., see TgRH88 Chromosome IV; Figure 2B). Given that
213 long read sequencing technologies improve sequence contiguity across tandem repeats
214 compared to shorter read technologies like Illumina (especially at tandem repeats that are
215 longer than the typical Illumina read), we mapped all tandem repeats with periods greater than
216 500 bp as well as select known repeats in the RH assemblies and mapped them along with
217 nucmer-generated pairwise sequence alignments for chromosomes IV and VI (Figure 2C). For
218 chromosome IV, the most striking difference between the long read assemblies and that
219 generated using Illumina is the expanded assembly at the locus harboring the well-
220 characterized 529 bp repeat (Figure 2C, top). We also found that loci harboring select tandem
221 repeat genes (e.g., TGME49_318130 and TGME49_319090) were also expanded in the long-
222 read *T. gondii* chromosome IV assemblies compared to that generated using solely Illumina
223 sequences. Alignments for *T. gondii* chromosome VI are also shown in figure 2C, in this case
224 between just the long read assemblies to illustrate differences and similarities. In multiple cases
225 the subtelomeric repeat sequences SAT350 and TGR4 were found in larger arrays in the
226 Nanopore assembly, as was the case for the known tandem rhoptry gene array encoding
227 ROP38. In contrast gene TGME49_240310 was found in a larger array in the PacBio assembly.

228 Overall these data indicate the utility of long read sequencing for resolving complex tandem
229 repeats in *T. gondii*, especially those that are larger than typical 1st or 2nd generation sequence
230 reads. Moreover with a few minor exceptions both existing long-read sequencing technologies
231 give similar results for the size and complexity of these repeats, and results in remarkably
232 similar estimates for nuclear chromosome and plastid genome sizes.



	*Tg*RH88 Nanopore Hybrid	TgRH88 PacBio Hybrid	TgRH Illumina
# contigs	23[1]	203[2]	2,648[3]
Total bases (bp)	64,918,878	67,055,676	69,349,265
Maximum contig length (bp)	12,055,564	12,055,564	5,672,850
Contig length N50 (bp)	6,778,623	6,690,982	64,945
Plastid genome length(s)[4]	34,855	35,031	29,462
Mitochondrial genome length(s)[4]	16,985; 15,838	Unknown	Unknown

[1]20 contigs, 1 scaffold; [2]193 contigs, 4 scaffolds containing 6 contigs total. [3]Contigs arrayed into 441 Scaffolds; [4]Based on homology searches using *T. gondii* mitochondrial or plastid genome sequences.

233

234 ***Figure 2: Comparisons between the current Nanopore assembly of T. gondii strain RH88***
235 ***and an existing long read (using PacBio RS and Illumina technology) and short read***
236 ***(Illumina only) assemblies.*** *A) Assembly statistics for each. B) Circos plot of Nucmer pairwise*
237 *alignments across all 3 assemblies for T. gondii chromosome IV. All alignments > 10000 bp*
238 *and > 90% identity are shown. C) Pairwise alignments for chromosomes IV (top) and VI*
239 *(bottom) along with locations of select tandem repeats identified either de novo (orange or blue*
240 *bars on the chromosome scaffolds) or known from prior studies (above or below chromosome*
241 *scaffolds). For T. gondii chromosome IV comparisons to both the PacBio and Illumina*
242 *assemblies are shown, while only the long-read comparison is shown for T. gondii chromosome*
243 *VI.*
244

245 **Nanopore read quality assessment and *de novo* genome assembly statistics**

246 To determine whether the chr VIIb/VIII fusion was unique to TgRH88 or was also present in
247 other isolates, we sequenced and assembled genomes of *Tg*ME49, *Tg*CTG, *Tg*ME49×*Tg*CTG
248 F1 progeny (CL13, S27, S21, S26, and D3X1), and *N. caninum* Liverpool strain (**Table 1**) by
249 performing 6 MinION runs with 6 R9.4.1 flow cells. On average, a single flow cell yielded
250 approximately 600,000 reads containing more than 7.8 Gb of sequences over a 48-hour run,
251 and for each strain, total yield varied from 0.6 Gb to 7.4 Gb (**Table S2**). The read length N50s
252 (the sequence length of the shortest read at half of the total bases) were longer than 18 Kb, and
253 the maximum read lengths varied between 116 Kb and 266 Kb (**Table S2**). The average Phred
254 quality scores (a measure of the quality of the base-call, representing the estimated probability
255 of an error (Ewing and Green 1998; Ewing *et al.* 1998)) for all libraries were ~10.0, except for
256 the *Tg*RH88 library, whose average Phred score was 8.9 (**Table S2**). Aligning the reads against
257 their most relevant "Reference" genome in ToxoDB (based on species and then closest
258 genotype) revealed that 96% of the *T. gondii* reads and 85% of the *N. caninum* reads could be
259 mapped, with a mean percent identity of ~86% (**Table S2**). After reads were corrected using
260 error-correction in Canu, 99% of the *T. gondii* reads and 98.07% of the *N. caninum* reads could
261 be mapped to their reference genomes, and the mean percent identities were ~95% (**Table S3**).
262 This first round of error correction (based on alignment and overlap between Nanopore reads as
263 implemented in Canu; (Koren *et al.* 2017)) was sufficient for making conclusions about
264 previously unappreciated structural variation within and between species.

265

266 The primary assemblies had a median number of contigs of 38.5 for the *T. gondii* strains
267 and 58 for the *N. caninum* Liverpool strain (**Table S4**). The median assembly size of the *T.*
268 *gondii* strains was 64 Mb, with a median contig length N50 of 6.63 Mb, and a median L50 of 4
269 (**Table S4**). The *N. caninum* primary assembly had a size of 61.8 Mb, with a contig N50 size of
270 6.38 Mb and an L50 of 4 (**Table S4**). Depending on the strain, contigs containing plastid

271 genome sequence ranged in size from 8 Kb to 214 Kb (Table 2), which is mostly inconsistent
272 with the known size of the plastid genome. While for the 8Kb contig (from strain S26) was likely
273 due to low sequence coverage, the larger contigs may be either concatenated versions of the
274 plastid genome or hybrid assemblies of plastid sequence along with some other non-nuclear
275 sequence. This suggests that the plastid genome may be more susceptible than the nuclear
276 genome to artificial concatenations during assembly of error-prone long sequence reads,
277 although we have not investigated this directly. In contrast to the organellar genomes the
278 nuclear *T. gondii* genomes assembled with high consistency across multiple strain types and
279 their F1 progeny. Although the mean percent identity of the aligned Canu-corrected reads to the
280 reference genome (~95%) was improved compared to the raw reads (~86%), we used a second
281 round of error correction to improve sequence accuracy. To do this, we mapped whole-genome
282 Illumina paired-end reads (SRA: SRR5123638, SRR2068653, SRR5643140, or ERR701181) to
283 the *Tg*RH88, *Tg*ME49, *Tg*CTG, or *Nc*Liv primary assemblies generated by Canu, and iteratively
284 polished the assembly contigs four times using Pilon v1.23. The polished contigs were then
285 reassembled using Flye v2.5. The resulting contigs/scaffolds were then assigned, ordered, and
286 oriented to chromosomes using relevant ToxoDB-48 reference genomes (**Table 2 and Table**
287 **S1**). These polished assemblies have been deposited in Genbank (under BioProject number
288 PRJNA638608).
289
290
291 The final assemblies of *Tg*RH88, *Tg*ME49, or *Tg*CTG consisted of 13 chromosome
292 contigs/scaffolds and varying numbers of unplaced fragments, with an average total size of
293 ~64.8 Mb (**Table 2**). The polished *N. caninum* Liverpool assembly was composed of 58 contigs,
294 showing a cumulative size of 62.1 Mb (**Table 2**). For all species and strains, the long reads and
295 high coverage led to highly improved contiguity of our assemblies compared to the reference
296 genomes. As reported in **Table 2**, with one exception (IIxIII F1 progeny S26), all the *T. gondii*

297 final assemblies were composed of 23-59 contigs, representing a 43-109-fold reduction in the
298 number of contigs in comparison to the ToxoDB-48_*Tg*ME49 assembly (2511 contigs; **Table 2**).
299 We performed one-to-one mappings between our Nanopore chromosome-sized contigs and all
300 aligning contigs from the ToxoDB-48_TgME49 assembly (Figure S1A) using Minimap2 and
301 found that the most significant contribution to overall genome structure was the elimination of
302 nearly all of the breaks within the ToxoDB-48_TgME49 scaffolds that are indicated within that
303 assembly as strings of N's at least 100 bp long (Figure S1A). We also compared the sequences
304 and to identify new sequence uniquely incorporated into the Nanopore chromosomes and
305 identified 42 regions of at least 10,000 bp ranging in size from 10,176 to 49718 bp (summing to
306 926 kb in total; yellow boxes in Figures S1A).
307 For *N. caninum*, compared to the existing assembly of the *Nc*Liv genome (Reid *et al.* 2012),
308 the total number of contigs in the *Nc*Liv final assembly was reduced from 247 to 58 (**Table 2**).
309 The contiguity improvement of long-read assembly was also evident by high contig length N50
310 values and low L50 values in these assemblies (**Table 2**). Furthermore, for all the long-read
311 assemblies, 55.2% of all chromosomes were fully assembled in single contigs and flanked by
312 two telomeric repeats with proper orientation (e.g., one end 5'-3' TTTAGGG and the other 5'-3'
313 CCCTAAA; **Table S5**). This is in contrast to version 48 of the *T. gondii* ME49 genome housed at
314 ToxoDB where 3 of the 14 chromosome assemblies have telomeric repeats on both ends, 7
315 have a telomeric repeat on one end, and the remaining 4 putative chromosomes have no
316 telomeric repeats at all. Upon aligning our assemblies to their respective reference genomes
317 (**Table S1**), we found that over 95.9% of the *Tg*RH88, *Tg*ME49, or *Tg*CTG final assembly
318 sequences could be mapped to their relevant reference genome. The alignments revealed that
319 our *de novo* assemblies had fewer mismatches and indels per 100 Kb and higher longest
320 alignment and total aligned length when compared to the assemblies housed on ToxoDB (**Table**
321 **3**). To assess genome assembly completeness, we used BUSCO analysis on the polished
322 *Tg*RH88, *Tg*ME49, and *Tg*CTG assemblies and compared the results to a similar analysis for

323 the unpolished assemblies. This analysis, which counts the number of single-copy orthologs
324 unambiguously identified in a genome assembly, found that for 215 such loci 88% of the
325 complete genes were recovered from the polished, long read based assemblies, while 23.3 -
326 54% were identified in the unpolished assemblies (**Table 3**). These data show the utility of our
327 sequence polishing for improving the accuracy of the assembly.
328

329 **Long-read assembly identifies 13 chromosomes in multiple *T. gondii* genomes**

330 To assess the structural correctness of all of the the long-read assemblies, we aligned our
331 *T. gondii* assembly sequences to the reference genome using Nucmer. As can be seen in
332 **Figure S1B**, all of the *T. gondii* long-read assemblies exhibited strong collinearity with their
333 corresponding reference genomes, barring a small number of putative inversions. Consistent
334 with our finding for *Tg*RH88 primary assembly, the chr VIIb/VIII fusion was observed in each of
335 our *T. gondii* assemblies (*Tg*ME49 is represented in **Figure 3A**, red box; and **Figure S1A,B**). To
336 further confirm this observation, we aligned our *Tg*ME49 corrected reads back against the
337 *Tg*ME49 long-read assembly and found an average read depth of 40× for the entirety of
338 TgME49_tig00000001_chrVIII, and 37× for the "breakpoint" (TgME49_tig00000001_chrVIII:
339 5090422 bp) of chrVIIb and chrVIII, indicating that this was unlikely due to an assembly error
340 (**Figure 3B**). We then mapped the corrected nanopore reads again to the ToxoDB-48_*Tg*ME49
341 reference genome, and the alignments showed that all of the reads that were mapped either to
342 the end of chrVIIb or to the beginning of chrVIII spanned the gap between the two
343 chromosomes, with an average coverage of 105× (**Figure 3C**). This link between reads aligning
344 to chromosome ends was not present in any other chromosome pair (for instance, between
345 chromosomes IX and X; **Figure 3D**).

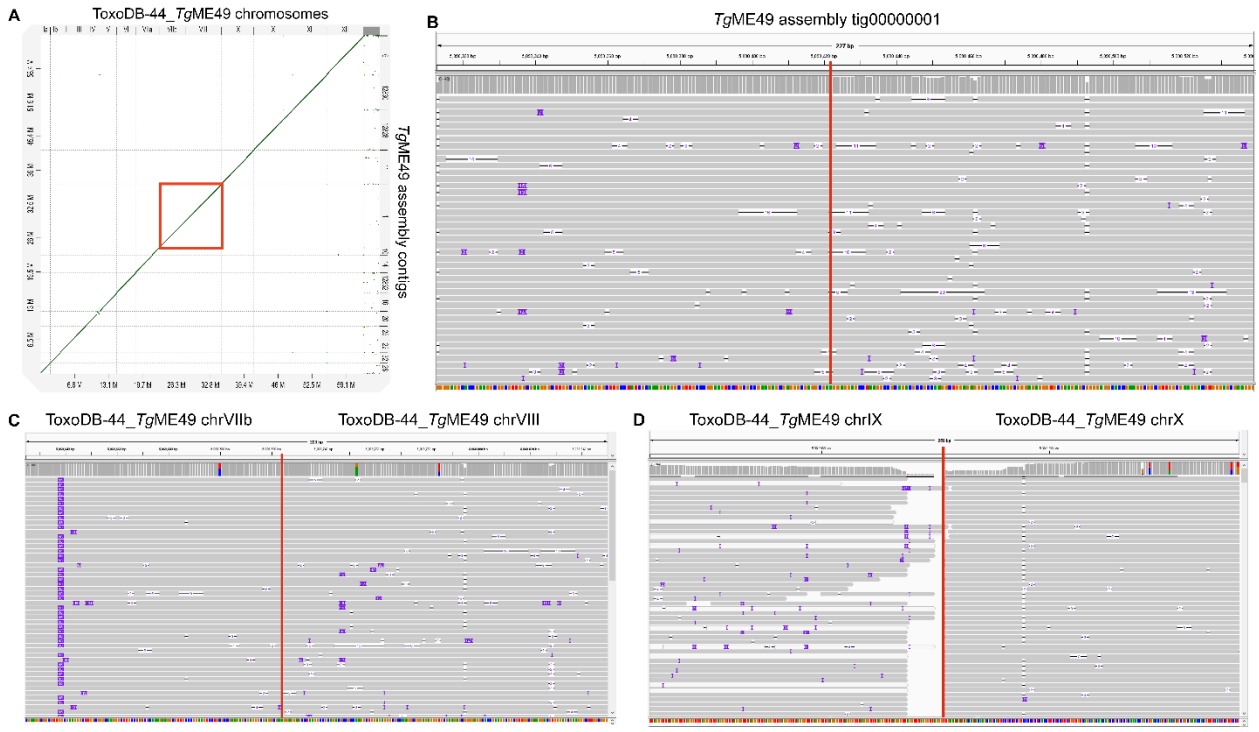
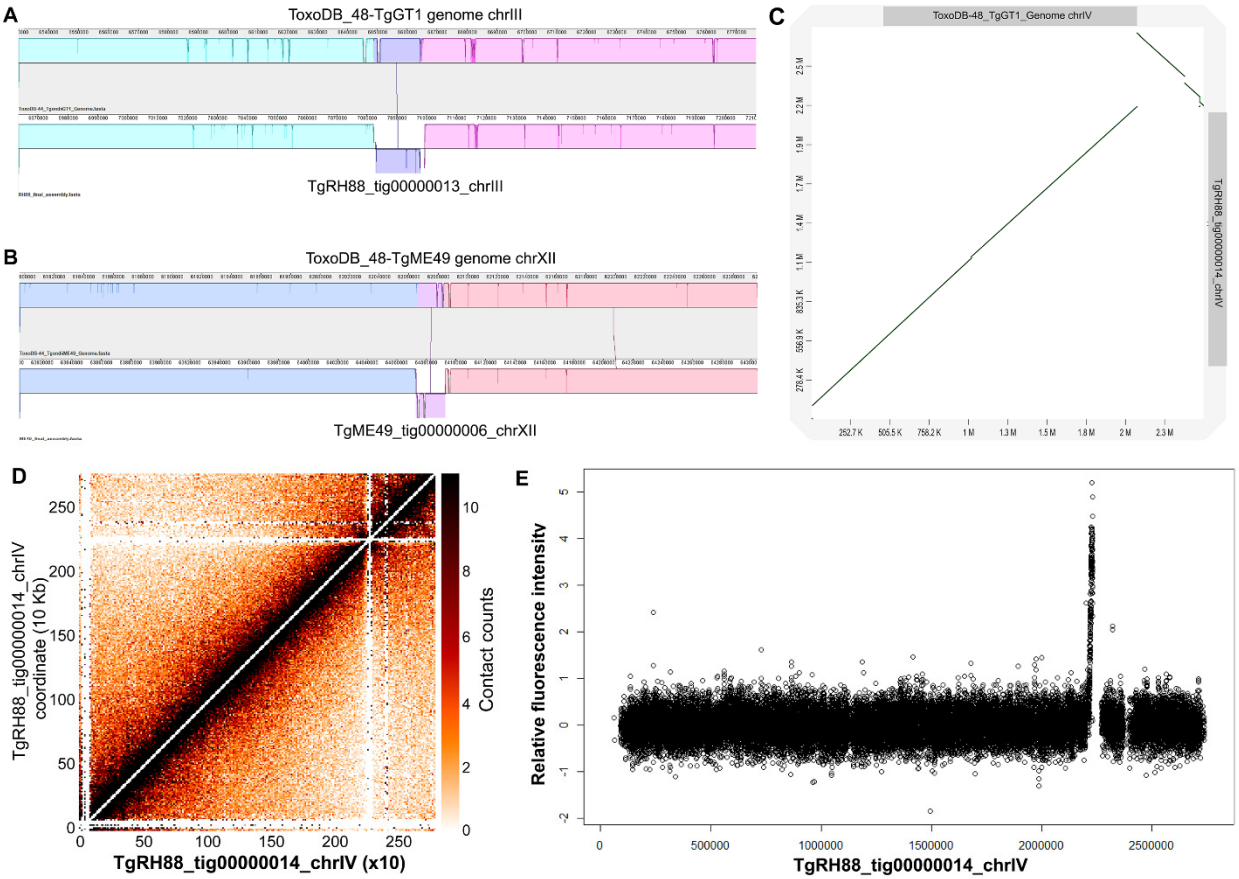


Figure 3. Long-read assembly identifies 13 chromosomes in the T. gondii genome from multiple strains. (A) *Dot plot showing the comparison of the TgME49 long-read assembly and the ToxoDB-48_TgME49 genome. Red box shows that the chromosomes VIIb and VIII in the ToxoDB-48_TgME49 genome are fused in a single contig, TgME49_tig00000001_chrVIII, in the TgME49 long read assembly. (B) Coverage of the "breakpoint" (TgME49_tig00000001_chrVIII: 5090422 bp, indicated by a vertical red line) of chromosomes VIIb and VIII with 37 Nanopore reads in the TgME49 long-read assembly. (C) Coverage of the edges (indicated by a vertical red line) of chromosomes VIIb and VIII with 105 Nanopore reads mapped to the ToxoDB-48_TgME49 genome. (D) Nanopore reads mapping to the end of chromosomes IX and X in the ToxoDB-48-TgME49 genome assembly, showing that Nanopore reads only map to the end of each chromosome and do not span the junction between these chromosomes (indicated by a vertical red line).*

This observation of a fusion between chrVIIb and chrVIII was in agreement with the observations in our TgRH88 assembly described above (**Figure 1C-E**). Similarly, when we mapped Hi-C data from (Bunnik *et al.* 2019) was aligned to our *Tg*ME49 long read assembly, the resulting interchromosomal contact-count map revealed 13 chromosomes in the *Tg*ME49 long read assembly by showing that each chromosome exhibited a single centromeric interaction with each other chromosome, and that TgME49_tig00000001_chrVIII was a

368 complete single chromosome (**Figure S2**). The intrachromosomal contact-count map of
369 TgME49_tig00000001_chrVIII showed a strong and broad diagonal and no aberrant signal
370 along the contig or at the the "breakpoint" of chrVIIb and chrVIII (**Figure S2B**). Similar patterns
371 were also observed in our *Tg*CTG, S27, and S21 assemblies (**Figure S2B**). Collectively, these
372 data show the *T. gondii* karyotype has been incorrectly calculated and contains 13, rather than
373 14, chromosomes. We refer to this fused chromosome as chrVIII in our assembly and have
374 eliminated chrVIIb.
375 The *Tg*CTG assembly had chromosome scale resolution, where 13 contiguous sequences
376 (contigs) corresponded to the 13 chromosomes. However, chromosome VIIa in the *Tg*ME49
377 assembly and chromosome XI in the TgRH88 assembly was spread across two contigs. Hi-C
378 data has historically been used to improve genome assemblies on the basis of contact
379 frequency depending strongly on one-dimensional distance (Dudchenko *et al.* 2017). That is, Hi-
380 C alignment to contigs in the correct order and orientation would reveal the canonical
381 intrachromosomal pattern of enriched interactions along the diagonal (where one-dimensional
382 genomic distance between bins is the smallest). Hi-C alignment to contigs in the incorrect order
383 and orientation would be illustrated by patterns associated with large-scale inversions, where
384 there is a high interaction frequency between bins placed far away from each other. Leveraging
385 this, we were able to determine the order and orientation of the contigs for chromosomes
386 ChrVIIa and XI in ME49 and RH88, respectively. These changes were incorporated prior to
387 submission of the assemblies to Genbank.
388
389 **Long-read assembly detects structural rearrangements in the *T. gondii* genome**
390 In addition to repeated elements, our assembly was capable of detecting gross structural
391 rearrangements (insertions, deletions, inversions, relocations, and translocations) since the
392 assembly based on long reads was highly contiguous and had near chromosome scale
393 resolution. We aligned the final assemblies of the 8 *T. gondii* strains to the reference genomes

394 and searched for structural variants using MUMmer (**Figure S1B**). Consistent with the data
395 shown in **Figure S1B**, most contigs in the long-read assemblies were collinear with the
396 chromosomes in the ToxoDB-48 genomes, but not always in a 1:1 correspondence.
397 Interestingly, we observed a 15.7 Kb inversion on chrIII in our *Tg*RH88 assembly, which was
398 absent in *Tg*ME49, *Tg*CTG, or any F1 progeny assembly (**Figure 4A**). We also detected
399 another ~20 Kb inversion on chrXII, which was present in *Tg*ME49, *Tg*CTG, and the F1 progeny
400 assembly, but not in the *Tg*RH88 assembly (*Tg*ME49 is represented in **Figure 4B**).



401

402 ***Figure 4. Long-read assembly reveals previously unknown inversions***
403 ***and the centromere location on chrIV in T. gondii.*** *(A) Inversion in the*
404 *RH88 long-read assembly on chromosome III relative to the ToxoDB-48-*
405 *TgGT1 assembly. (B) Inversion in the ME49 long-read assembly on*
406 *chromosome XII relative to the ToxoDB_44-TgME49 genome. (C) Dot plot*
407 *comparison of the TgRH88 long-read assembly and the ToxoDB-48_TgGT1*
408 *genome showing a 429.3-Kb inversion at 2,096,529-2,525,795 bp on chrIV.*
409 *(D) Intrachromosomal Hi-C contact-count heat map plotted using the*
410 *sequence of contig_14 in TgRH88 long-read assembly showing a clear*

411 *centromere signal at position 2.2-2.3 Mb. (E) ChIP-on-chip signal of*
412 *centromeric histone 3 variant (CenH3) (Brooks et al. 2011) plotted using the*
413 *TgRH88 long-read assembly as coordinate.*

414 Centromeres for 12 of the 13 *T. gondii* chromosomes have been identified using ChIP-on-
415 chip (chromatin immunoprecipitation coupled with DNA microarrays) of centromeric and peri-
416 centromeric proteins, but locations of centromere of chrVIIb and chrIV remain unknown (Brooks
417 *et al.* 2011; Gissot *et al.* 2012). For chrVIIb, no hybridization of the centromeric probe to the
418 genomic chip was detected in the ChIP-on-chip assay (Brooks *et al.* 2011), which could be
419 explained by our observation that chrVIIb and chrVIII are a single chromosome, and the
420 centromere of this large chromosome appears to be in the center of this "fused" chromosome, in
421 an area that was previously thought to be the beginning of chromosome VIII (**Figure 1E**). For
422 chromosome IV, two inconsecutive peaks of hybridization were detected at positions 2,501,171-
423 2,527,417 bp on chrIV and 1-9968 bp in the unplaced contig AAQM03000753 in the *Tg*GT1
424 genome based on published ChIP-on-chip data (Brooks *et al.* 2011). Our *Tg*RH88 assembly
425 successfully relocated the sequences in AAQM03000753 into chrIV and revealed a 430.9-Kb
426 inversion event at 2,096,529-2,527,423 bp on chrIV relative to the ToxoDB GT1 reference
427 genome (**Figure 4B**). This inversion was unlikely to be due to an assembly error since it was
428 shown in all of our *T. gondii* assemblies, and was supported by alignment of 128 canu-corrected
429 reads spanning the boundaries of the *Tg*RH88 assembly. When we mapped the ChIP-on-chip
430 data obtained from (Brooks *et al.* 2011) to our *Tg*RH88 assembly (after remapping the probe
431 sequences to our TgRH88 assembly), we resolved the chrIV centromere to a single significant
432 signal peak at 2.20-2.23 Mb on chrIV (**Figure 4C** and **Figure S3**). This finding was also
433 supported by published Hi-C data re-aligned to our Nanopore assemblies, since the
434 intrachromosomal contact count map showed a clear interchromosomal contact signal at 2.20-
435 2.30 Mb on chrIV in the *Tg*RH88 assembly (**Figure 4B**). Collectively, our data not only resolved
436 the molecular karyotype of *T. gondii*, but also resolved the precise location of the chrIV
437 centromere.

438

439 **Long-read assembly adds new sequences to the *T. gondii* reference genome**

440 As shown in **Figure 1C** and **Table S5**, each chromosome-sized contig of our long-read *de novo*
441 assemblies was longer than its cognate chromosome of the ToxoDB-48 *T. gondii* reference
442 genome, and for TgRH88, TgME49 and TgCTG we were able to assemble between 1.7 and 3.9
443 Mb of previously unlocated and/or unassembled sequence to the chromosomes of these
444 assemblies. These new sequences were scattered across the genome, and filled in nearly all of
445 the sequence gaps found in the reference genome. The new sequences added by the long-read
446 assemblies extended the subtelomeric regions of the chromosomes in the *T. gondii* reference
447 genome. While 4 chromosomes of the ToxoDB-48_*Tg*ME49 genome contained no telomeric
448 repeat and 7 were missing one of the telomeric repeats, all of the chromosome contigs in the
449 *Tg*ME49 long-read assembly were assembled up until both telomeric caps (**Table S5**). Both
450 telomeres were found in 12 out of the 13 chromosomes in the *Tg*CTG long-read assembly, and
451 one chromosome contig lacked one of the telomeric repeats, whereas only 5 chromosomes in
452 the ToxoDB-48_*Tg*VEG genome contained one telomere and no telomeric repeat was found in
453 the rest of the chromosomes (**Table S5**). Similarly, both telomeres in 7 chromosomes and one
454 telomere in 6 chromosomes were resolved in the *Tg*RH88 long-read assembly, while there were
455 only 2 chromosomes in ToxoDB-48_*Tg*GT1 genome that contained one telomere (**Table S5**).

456 In addition to telomeres, the bulk of the remaining new sequence was due to tandem arrays
457 of sequence (both coding and non-coding). For example, two repetitive gene sequences are
458 used for high sensitivity detection of *T. gondii* in tissue and environmental samples, the B1 gene
459 (Burg *et al.* 1989) and the so-called “529 bp repeat” (Reischl *et al.* 2003; Edvinsson *et al.* 2006).
460 The precise copy number for these genes has been impossible to determine using 1st and 2nd
461 generation sequencing technologies and/or molecular biological experiments like Southern
462 Blotting. Therefore, we used a curated blastn approach to quantify copy number for each of
463 these sequences across our respective Nanopore assemblies. As shown in **Figure 5A**, copy

464 number for the B1 gene was dramatically higher in our Nanopore assemblies compared to
465 existing ToxoDB assemblies (as expected). What was unexpected, however, was that copy
466 number for this gene was lower than that predicted in the literature, ranging between 9 and 19
467 tandem copies depending on the strain (**Figure 6A**), compared to quantitative blotting-based
468 estimates of 35 (Burg *et al.* 1989). Copy number at this locus was stable, in that for all of the
469 queried IIxIII F1 progeny copy number for each was identical to the parent from which it
470 obtained that chromosomal segment. In contrast to the B1 locus, the "529 bp repeat" locus
471 varied dramatically between isolates and these same F1 progeny. Copy number ranged from 85
472 to 205, and copy number at this locus for all F1 progeny varied independently of the underlying
473 genotype for that region (white letters and green/blue in **Figure 5B**). The size of this genome
474 expansion is best illustrated by the whole chromosome alignment shown in **Figure 5E**
475 comparing the *T. gondii* ME49 529 bp repeat locus in the version 48 assembly on ToxoDB to
476 our Nanopore assembly (**Figure 5E**). Importantly the 529 bp repeat locus occurs near a
477 sequence assembly gap, and our long read assembly closed this gap (see below and **Figure**
478 **5D**), giving the most accurate estimate of 529 bp repeat copy number in any *T. gondii* strain
479 which again varies compared to estimates in the literature (ranging from 200-300 copies; e.g.,
480 (Reischl *et al.* 2003; Edvinsson *et al.* 2006)). Regardless, similar to tandem gene arrays
481 discussed above in **Figure 6**, it appears that even noncoding repeats like the B1 gene and the
482 529 bp repeat can also change in number irrespective of whether there is sexual recombination.
483 While the 529 bp repeat and B1 gene are found in a single locus, other tandem repeats like
484 TgIRE and SAT350 are found across multiple subtelomeric genomic locations (Echeverria *et al.*
485 2000; Clemente *et al.* 2004). We mapped these sequences across the genome in all of our
486 sequenced *T. gondii* strains, and again found different results depending on the queried locus.
487 For the TgIRE sequence, a 1919 bp repeat, chromosome-wide copy number in our sequenced
488 strains was ~2x that found in the reference sequences (Figure 6C), but overall copy number
489 was similar across all of our Nanopore-derived sequences. The SAT350 sequence was also

found at much higher copy number in our Nanopore assembled chromosomes compared to the reference sequences (Figure 6D), but was more variable in the F1 progeny clones. This could be due to changes during sexual recombination as for the 529 bp repeat above, or differences in sequence coverage for the F1 progeny. Regardless, our approach has provided better resolution of copy number for 4 well-known tandem repeat sequences, and allowed us to determine which may be more susceptible to copy number change. Moreover, unappreciated strain differences in copy number at these loci may adversely affect interpretation of PCR-based detection assays, especially those using more quantitative methods.

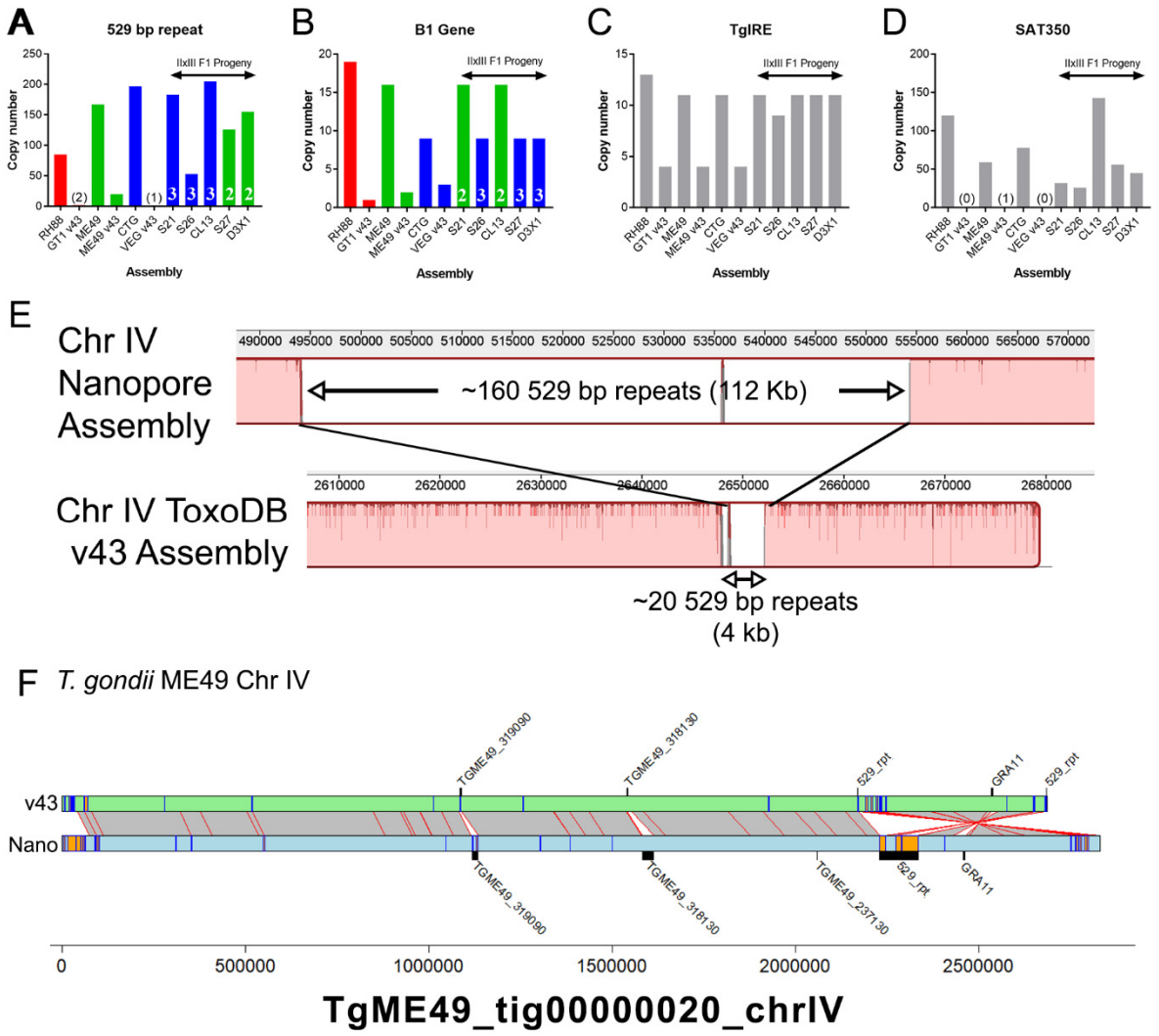


Figure 5: Long read sequence assemblies precisely resolve canonical repeat

500 *sequences and identify additional expansions at gene-harboring loci. (A-D)*
501 *Estimated copy number for Nanopore assemblies and existing genome*
502 *assemblies on ToxoDB ("v48") for T. gondii strain types 1, 2 and 3 and IIxIII F1*
503 *progeny. In all cases, Nanopore assemblies identified higher numbers of each*
504 *repeat locus. In the F1 progeny, B1 gene copy number tracked directly with*
505 *the genotype (type 2 or 3) at that locus (A), while these same F1 progeny*
506 *harbored unique numbers of 529 bp copies, all of which were not only distinct*
507 *from their respective genotypes of origin but distinct from one another (B). The*
508 *TgIRE and SAT350 repeats also were better resolved in our Nanopore*
509 *assemblies (C,D) although determining genotype of the corresponding region*
510 *is not possible since these repeats are found at multiple locations throughout*
511 *the genome. (E) Whole chromosome alignment focused on the 529 bp repeat*
512 *region for the v48 ToxoDB assembly (bottom) and our Nanopore-based*
513 *assembly (top). Expansion of the known genome sequence at this locus in the*
514 *Nanopore sequence compared to the ToxoDB assembly is clear, and*
515 *consistent with our identification of ~140 previously unknown 529 bp repeats*
516 *in the ME49 genome. (F) Alignment and annotation of repeat sequences of*
517 *ME49 v48 ToxoDB chromosome IV and that from our polished Nanopore*
518 *assembly. Grey bars with red borders indicate mapping regions ≥10,000 bp*
519 *determined using nucmer, while orange boxes with blue borders indicate*
520 *tandem repeats with period sizes ≥ 500 bp and at least 2 copies. Bars that*
521 *appear orange are larger than those that are only blue. Incorrect inversion on*
522 *the right arm of chromosome IV in the ToxoDB assembly is evident, as is the*
523 *more accurately resolve 529 bp repeat locus that was likely a cause for the*
524 *inversion in standard assemblies from multiple strains.*
525

526 With respect to chromosome IV, mapping the 529 bp repeat to our Nanopore assemblies
527 while comparing them to reference genomes in ToxoDB identified what appears to be an
528 inversion in the right arm of chromosome IV (Figure 6F). This inversion was present in the *T.*
529 *gondii* VEG and GT1 reference genomes, suggesting that it was due to systematic assembly
530 error. This inversion in the ToxoDB reference genome is flanked by multiple tandem repeats
531 (identified using tandem repeats finder; blue/orange boxes in **Figure 6F**). In addition, the 529 bp
532 repeat can be found at the end of each inversion ("529_rpt", **Figure 6F**) in the ToxoDB
533 chromosome, but in our Nanopore assembly the 529 repeat cluster is only found in one location
534 (where it is greatly expanded compared to those in the ToxoDB reference). These data provide
535 strong evidence that chromosome IV is incorrectly assembled in multiple ToxoDB reference
536 genomes due to misassembly of the 529 bp repeat locus, and our Nanopore assembly has
537 resolved this discrepancy in multiple *T. gondii* genomes.

538 **Long-read assembly resolves tandem duplicated locus structure in the *T. gondii* genome**

539 Many scaffold gaps within chromosomes remain unresolved (**Table 2**) in the *T. gondii*
540 reference genome particularly in the regions that contain repeated elements. Our long read
541 assembly closed nearly all of the gaps in the *T. gondii* and *N. caninum* genomes (Tables 2 and
542 S6). Furthermore, many unplaced sequences in the *T. gondii* reference genome were
543 assembled into contigs in our assembly. For instance, the unplaced contig KE140372 in the
544 ToxoDB-48_*Tg*ME49 genome, which was 2194 bp in length and contained a sequence
545 encoding a rhoptry protein 4 paralog, was assembled in TgME49_tig00000028_chrIa in our
546 *Tg*ME49 assembly. Unplaced contigs in the ToxoDB-48_*Tg*GT1 genome, AAQM03000823 and
547 AAQM03000824, were assembled in TgRH88_tig00000013_chrIII in our *Tg*RH88 assembly.

548 We have had a long-standing interest in variation at tandemly expanded gene clusters
549 and how this affects *T. gondii* virulence which was first spurred by our identification of ROP5
550 gene cluster as being a critical determinant of virulence in the mouse (Reese *et al.* 2011). Our
551 genome-wide analyses of copy number variation across multiple strains and species (Adomako-
552 Ankomah *et al.* 2014) identified 53 putative tandemly expanded gene clusters in *T. gondii*
553 (shown in Table S6), some of which (such as MAF1) have been shown to be important in host
554 pathogen interactions (Adomako-Ankomah *et al.* 2016). In that initial study we used sequence
555 coverage to infer copy number and we were eager to refine this analysis using the long-read de
556 novo assemblies presented here.

557 The gap closure enabled us to determine the number, order and orientation of *T. gondii*
558 duplicated loci including *ROP7*, *ROP5*, *ROP38*, *MIC17*, *MAF1*, and *TSEL8*. The *ROP7* locus is
559 represented in **Figure 6A**, where we identified two unresolved scaffold gaps on chrIa in the
560 ToxoDB-48_*Tg*ME49 genome (black bars), and these gaps marked the site of *ROP4/7* locus.
561 This entire region was spanned by a single contig, TgME49_tig00000028_chrIa, in our *Tg*ME49
562 assembly (the blue bar in **Figure 6A**). Aligning the *ROP7* genomic sequence (ToxoDB:
563 TGME49_295110) against TgME49_tig00000028_chrIa using BLASTN revealed 3 copies of the

564 *ROP7* repeat, while only one was predicted in the ToxoDB-48_*Tg*ME49 genome (**Figure 6A**
565 **and 6B**) Interestingly, while the ToxoDB-48_*Tg*ME49 genome identified one copy of the *ROP4*
566 gene (GenBank: EU047558.1), our *Tg*ME49 assembly showed that 2 copies of *ROP4* exist in
567 this locus, one of which was found between the first and the second copy of *ROP7* (**Figure 6B**).
568 To validate this finding, we identified 12 individual Canu-corrected reads that spanned this entire
569 tandem array, and each one that we examined provided evidence for 3 copies of *ROP7* and 2
570 copies of *ROP4* arranged in the order ROP4_1-ROP7_1-ROP4_2-ROP7_2-ROP7_3 (**Table**
571 **S6**). The copy number and copy order of other known tandem locus expansions (taken from the
572 supplementary table in (Adomako-Ankomah *et al.* 2014)) in the strains we sequenced were
573 identified and are listed in **Table S6**. Surprisingly, we observed changes in copy number at
574 some of these loci when we compared the parental (*Tg*ME49 or *Tg*CTG) and the progeny
575 (CL13, S27, S21, S26, and D3X1) assemblies (**Figure 6C**). For example, while the *ROP5* locus
576 had 9 copies in our *Tg*ME49 assembly and 4 in our *Tg*CTG assembly, it harbored 6 copies in
577 the F1 progeny S27 and S21, and 7 in S26 (**Figure 6C**). There was an array of 7 tandem copies
578 of *MIC17* in *Tg*ME49 and *Tg*CTG assemblies, whereas it was present in 6 copies in the S26
579 assembly (**Figure 6C**). These data indicated that changes in copy number and order at tandem
580 gene arrays can occur during sexual recombination. Moreover the ease with which a Nanopore
581 assembly can be generated and assembled provides a new way to assess the occurrence and
582 impact of acute changes in copy number that occur during asexual and sexual propagation of
583 parasites like *T. gondii*. Whole chromosome alignments are shown for chromosomes Ia, VI and
584 XII to further illustrate the increase in sequence size at the chromosome level for loci like ROP7,
585 ROP38 and ROP5, as well as other known tandem gene arrays (**Figure 6D, E and F**).
586 We performed a similar analysis genome-wide for the remaining previously identified
587 tandem expansions ((Adomako-Ankomah *et al.* 2014); **Table S7**), and were able to further
588 curate this list of putative repetitive loci. While in some cases our predictions from shotgun
589 sequence read coverage was similar to that predicted using Nanopore assembly (as for those

590 genes described in **Figure 6** as well as many others in **Table S7**), for some we were able to
591 determine that there was in fact no evidence for the presence of a tandem gene array at that
592 locus. For example expanded locus 9 (EL9; TGME49_319090), EL19 (TGME49_204560) and
593 EL27 (TGME49_264420) are all likely to be single copy genes based on our Nanopore
594 assemblies (**Table S7**), even though based on sequence coverage alone they ranged in
595 predicted copy number between 15 and 60 (**Table S7** and (Adomako-Ankomah *et al.* 2014)).
596 While the reasons for each of these discrepancies will be locus specific, it is likely that the
597 overestimation of copy number using sequence coverage in our prior study was due to the
598 presence of low complexity sequence (e.g., short tandem repeats) across the single copy gene
599 rather than being due to actual tandem expansion. Overall, our current assemblies refine and
600 further curate this locus list, providing the most accurate estimate to date of which loci encode
601 tandem gene arrays and how their copy number varies across strain and species.

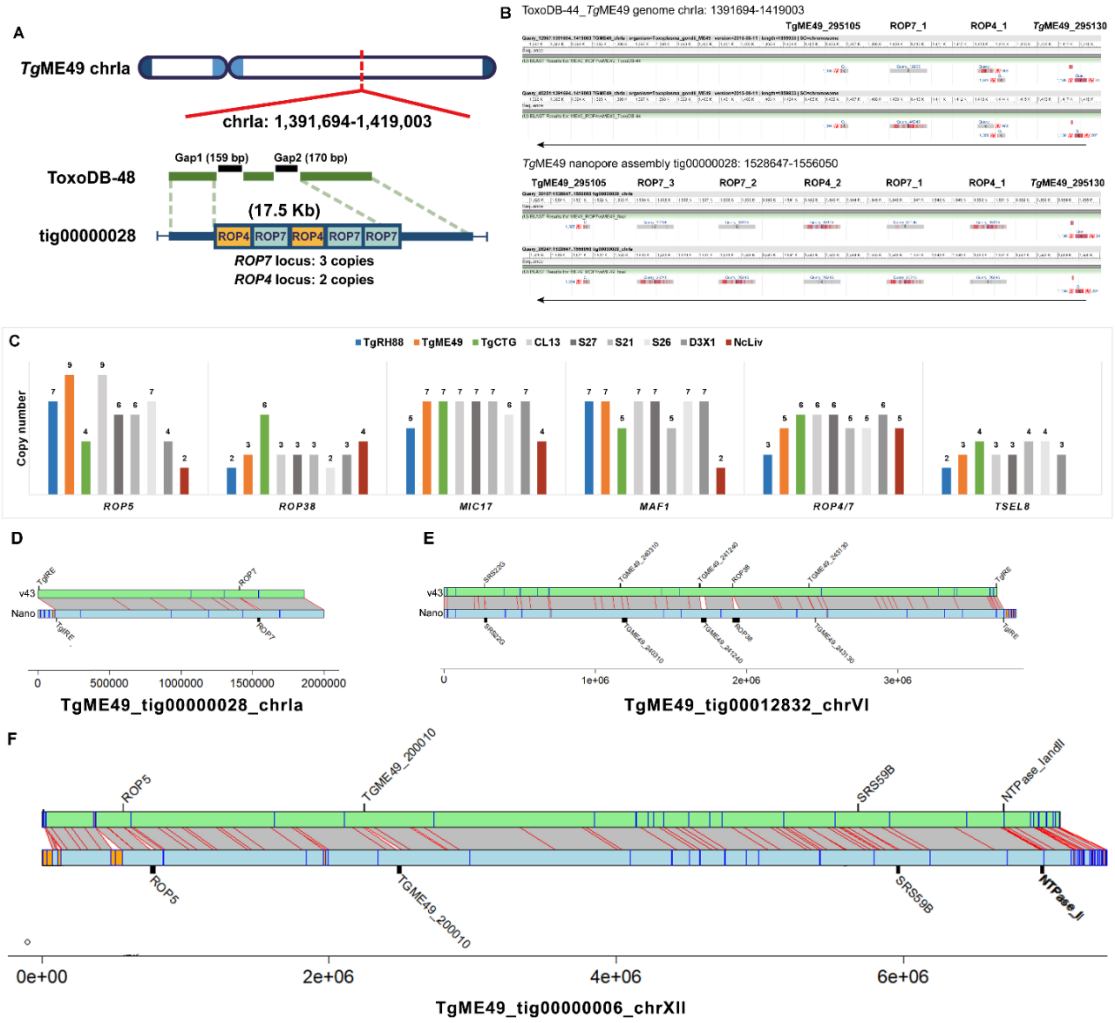


Figure 6. Long-read assembly resolves duplicated locus structure in T. gondii genome. *(A) Two unresolved scaffold gaps on chrIa in ToxoDB-48_TgME49 genome span a 17.5-Kb tandem repeat containing multiple copies of ROP4 and ROP7. The ROP4/7 gaps are closed by the TgME49 long-read assembly (TgME49_tig00000028), revealing a tandem array of 5 copies of this gene in the order shown. (B) BLASTN alignment of the ROP4/ROP7 coding sequence in the ToxoDB-48_TgME49 genome (upper panel) and the TgME49 long-read assembly (lower panel). (C) Copy number determination at 6 canonical tandem gene arrays across 8 T. gondii strains and 1 N. caninum strain. Data from CL13, S27, S21 and S26 show that copy number can change during sexual recombination since copy number in these F1 progeny clones do not match copy number in either parent. (D-F) Whole chromosome alignments between ME49ToxoDB-48 and our Nanopore assemblies at loci harboring tandem gene arrays. Grey boxes with red borders indicate 1-to-1 mapping regions ≥ 10,000 bp determined by nucmer and orange/blue boxes are as described in Figure 5. Black bars indicate size of select tandem repeats in the ToxoDB and Nanopore assemblies.*

620

621 As described above all of our sequence assemblies increased the size of the contiguous
622 assemblies by 1-3 Mb. While some of this new sequence is most certainly derived from gene-
623 poor regions containing simple tandem repeats (and some of these regions are annotated as
624 yellow boxes in **Figure S2**), the relatively high gene density of the *T. gondii* genome led us to
625 hypothesize that some of this "new" sequence should be derived from gene sequences that
626 were previously masked by assembly gaps. Therefore we used BLASTN and custom parsing
627 scripts to identify genome expansions in our Nanopore assembly relative to the v48 sequence
628 on ToxoDB specifically using all available predicted genes as query sequences. Overall, we
629 identified 62 gene-containing loci that were at least 10 Kb larger in our Nanopore assembly
630 compared to ToxoDB v48, representing 1.2 Mb of sequence. These expansions are represented
631 in **Figure S4** as yellow blocks, and are shown along with known tandem gene arrays (red
632 blocks) and existing sequence gaps (black lines). Well known tandem gene arrays that are
633 collapsed in 1st and 2nd generation sequence-based assemblies like MAF1 and ROP5 were
634 identified in this analysis (**Figure S4**), confirming the accuracy of the approach. What was
635 unexpected was the unequal distribution of these "expansions" in our genome-wide analysis
636 (e.g., compare chromosomes XI and X). These expansions are due to increased copy number
637 as well as previously unknown insertions of repetitive sequence. In addition to these gene-
638 containing loci, we have identified all tandem repeats with a period size > 500 bp in our de novo
639 assembly of *T. gondii* ME49 and estimated their copy number in our de novo assemblies of *T.*
640 *gondii* RH88 and CTG (Table S8).

641

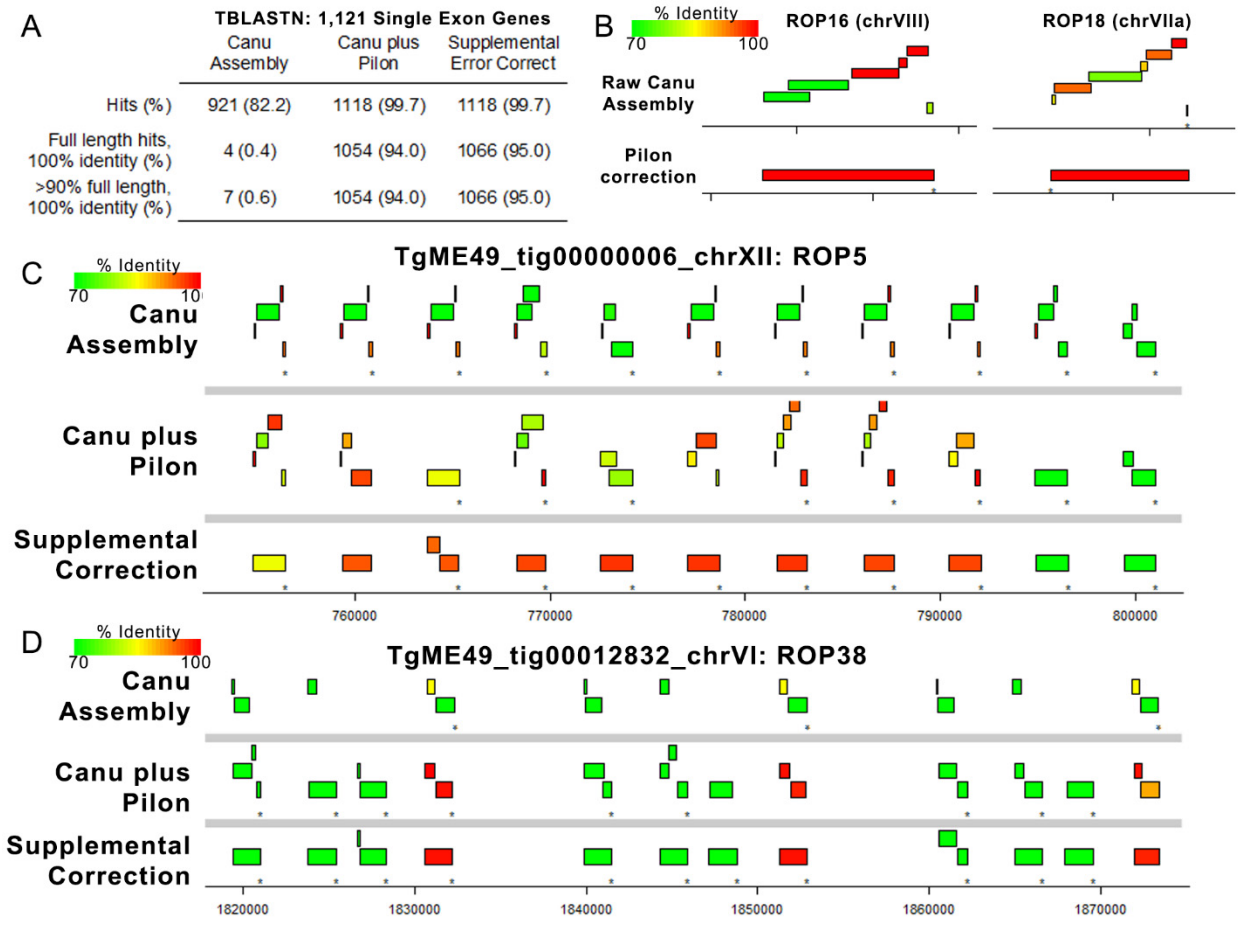
642 **Standard error correction methods for tandem gene arrays fail to remove extensive**
643 **homopolymeric repeats that lead to artifactual pseudogenization**

644 Error correction using Pilon is a common practice after generation of single molecule
645 long-read assemblies since the overlap-based correction employed by assemblers like Canu

646 fails to correct systematic errors in the data. For Nanopore, homopolymer runs as short as 3 bp
647 are often truncated in the consensus [need citation], leading to extensive artificial gene
648 pseudogenization. As described above hybrid assembly or error correction can eliminate many
649 of these systematic errors leading to improved protein coding gene annotations (Table 3). To
650 examine this is greater detail we used TBLASTN to align *T. gondii* ME49 (version 48) protein
651 sequences for all single exon genes to the ToxoDB reference genome, the primary Canu
652 assembly, and to our hybrid-corrected assembly. As shown in Figure 2A the raw Canu
653 assembly had only 4 full length, identical sequence hits to single exon genes, while the Pilon-
654 corrected assembly predicted nearly all of the query single exon genes with 100% identity and
655 100% sequence coverage (1054/1121; 94%; Figure 7A). The effectiveness of Pilon error
656 correction is shown for 2 single exon, single copy genes (ROP16 and ROP18; Figure 8B),
657 where the coding sequence had numerous frameshifts causing fragmented mapping in the Raw
658 assembly but resolved to a single gene with 100% identity and sequence coverage after Pilon
659 correction. In contrast to these single copy genes, we found that Pilon-based error correction
660 performed much more poorly at multicopy loci like those encoding ROP5 and ROP38. As
661 shown in **Figure 7C** and **D**, many of the predicted ROP38 and ROP5 coding sequences are still
662 highly fragmented after Pilon error correction, presumably due to an inability of Pilon to assign
663 enough reads to each copy to correct what are mostly homopolymeric repeat errors. We wrote a
664 custom Perl script to correct remaining homopolymer errors in tandem gene arrays (using
665 Illumina sequence read alignments to either extend homopolymers up to 10 bp or truncate by a
666 single base pair) and found that this eliminated many of the artefactual pseudogenes for the
667 tandem gene expansions at the ROP5 and ROP38 loci (see bottom of **Figure 7C** and **D**). Note
668 that unlike single copy genes where Canu plus Pilon correction was sufficient to correct them
669 (**Figure 7B**), we only eliminated these likely artefactual pseudogenes after running our
670 supplemental correction scripts (see Materials and Methods for specific details about the
671 correction script). We have deposited these corrected sequences in Genbank separately from

672 our genome assemblies for all the loci characterized in Figure 6C (submission numbers

673 pending).



675 *Figure 7: Error polishing with Pilon is effective for most single copy genes but*
676 *resolution of tandem gene expansion errors requires supplemental correction. (A)*
677 *We identified 1121 single exon genes with predicted proteins that mapped with*
678 *100% identity and 100% coverage using TBLASTN against the ToxoDB-48*
679 *ME49 genome. Then, we mapped these against the raw assembly generated by*
680 *Canu, the polished assembly generated by 4 rounds of Pilon, and after 4 rounds*
681 *of supplemental error correction. Pilon error correction was sufficient for perfect*
682 *mapping of 94% of the query single exon genes (compared to only 0.4% for the*
683 *raw Canu assembly), and supplemental error correction only increased this*
684 *mapping percentage slightly. (B) Plots representing TBLASTN analysis of*
685 *protein sequences from two single copy genes showing the improved mapping*
686 *achieved by Pilon-based error correction. Mapping identity is indicated by the*
687 *color of the box representing the alignment. (C-D) Plots representing protein*
688 *sequence from the ROP5A (C) or ROP38 (D) gene mapped using TBLASTN*
689 *against the raw Canu-only assembly, the Pilon-corrected assembly and the*
690 *region corrected using our supplemental approach tailored to tandem gene*
691 *arrays. Both loci have multiple pseudogenes in the Canu-only and Canu-plus*
692 *Pilon assemblies, but many of these errors are removed upon supplemental*

correction. The presence of a pseudogene in the ME49 ROP5 locus has been predicted before based on direct sequencing, suggesting that this may represent the most accurate version of the ME49 ROP5 locus sequenced to date.

Long-read assembly revises *N. caninum* karyotype its lack of synteny with *T. gondii*

The comparison of *Tg*RH88 long-read assembly and annotation with the ToxoDB-48_*Tg*GT1 genome revealed a high level of collinearity between the two genomes with no large-scale rearrangement between chromosomes (except for the chrVIIb/VIII fusion) (**Figure 8A**), whereas a large number of chromosomal translocations and inversions were observed in the *Nc*Liv long-read assembly with respect to ENA_*Nc*Liv genome (**Figure 8B**). For instance, the NcLiv_tig0000052 in *Nc*Liv long-read assembly contained a portion of chrIX and a portion of chrX in the current *N. caninum* genome (**Figure 8B**). A portion of NcLiv_tig00000006 in *Nc*Liv long-read assembly was mapped to the current *N. caninum* chrX, while the remainder of NcLiv_tig00000006 was mapped to chrIX and a region of it was inverted (**Figure 8B**). In addition to this, some chromosomal regions in the *Nc*Liv long-read assembly did not show any synteny with the ENA_*Nc*Liv genome (**Figure 8B**). Similar chromosomal rearrangement patterns were observed in other *N. caninum* strains (as described in the co-submitted paper).

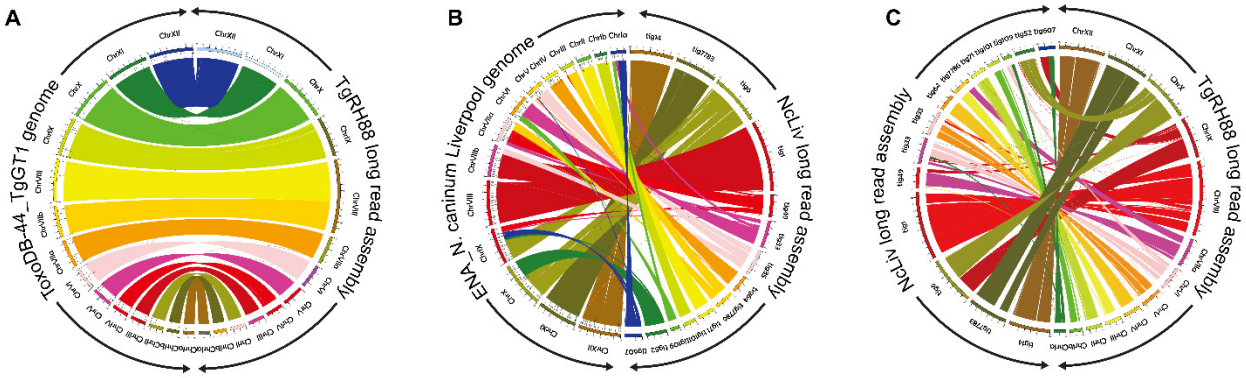


Figure 8. Long-read assembly reveals N. caninum karyotype and its synteny with T. gondii. (A) Circos plot showing high synteny between the TgRH88 long-read assembly and the ToxoDB-48_TgGT1 genome. (B) Circos plot showing the chromosomal translocations and inversions in NcLiv long-read assembly compared to the ENA_NcLiv genome. (C) Circos plot showing the syntenic relationship between TgRH88 and NcLiv long-read assembly.

717

718 While previous studies (Reid *et al.* 2012; Lorenzi *et al.* 2016a) showed that the current *T.*
719 *gondii* and *N. caninum* reference genomes were highly syntenic, the comparison of *Tg*RH88
720 long-read assembly with the *Nc*Liv long-read assembly showed smaller blocks of synteny in
721 most regions between the two genomes (**Figure 8C**). Many individual contigs of the *Nc*Liv long-
722 read assembly were shown to be mapped to multiple chromosome contigs in the *Tg*RH88 long-
723 read assembly. For example, a 3.26-Mb region and a 4.43-Mb region of TgRH88_tig00000006
724 of the *Nc*Liv long-read assembly were in synteny with regions on TgRH88_tig00000001 and
725 TgRH88_tig00000003 of *Tg*RH88 long-read assembly, respectively (**Figure 8C**). Moreover,
726 some regions of *Nc*Liv long-read assembly (e.g. NcLiv_tig00007783: 2,000,000-2,500,000 bp)
727 exhibit no synteny with the chromosomes of *Tg*RH88 genome (**Figure 8C**). Overall, our long-
728 read assembly revealed a new and accurate *N. caninum* karyotype, and revised the genomic
729 synteny between the two closely-related species, *T. gondii* and *N. caninum*.

730

731 **DISCUSSION**

732 The first wave of genome sequencing using first- and second-generation sequencing
733 technologies revolutionized our ability to link phenotype to genotype in diverse strains of
734 *Toxoplasma gondii* and its near relatives. Sequence from multiple isolates have been made
735 publicly available and hosted on ToxoDB.org and were outlined in a recent publication (Lorenzi
736 *et al.* 2016a). These genomes, while of great use, were expectedly incomplete due to the
737 presence of hundreds to thousands of sequence assembly gaps depending on sequencing
738 depth and methodology. Recent years have experienced the impact of so-called “3rd
739 generation” technologies that have revolutionized the speed, cost and efficacy of de novo
740 sequence assembly, and most importantly provide a means to dramatically improve existing
741 sequence assemblies. The critical feature of these approaches is the fact that they generate
742 sequence read lengths hundreds to thousands of times longer than those generated using 1st

743 and 2nd generation approaches, and can span repetitive sequence (both short tandem repeats
744 and larger segmental duplications). This technology provides a unique opportunity to greatly
745 improve the assembly of the *T. gondii* genome, particularly at repetitive loci which are known to
746 encode diversified secreted effectors (Adomako-Ankomah *et al.* 2014; Adomako-Ankomah *et al.*
747 2016).

748 Our work revises the *T. gondii* and *N. caninum* karyotypes by identifying a previously
749 unappreciated fusion between segments previously thought to be distinct chromosomes (VIIb
750 and VIII). Throughout the history of the *T. gondii* genome the number of linkage groups has
751 been a moving target and has become more precise as new mapping and sequencing
752 technologies have become available. Initial genetic mapping experiments and high molecular
753 weight Southern blotting identified 11 linkage groups (Sibley and Boothroyd 1992), while a
754 denser map in later studies identified 13 linkage groups (Su *et al.* 2002). This particular map
755 was still not fully representative of the *T. gondii* karyotype, as markers known to be on
756 chromosome VI were included in linkage group X, chromosome XII was split into 2 linkage
757 groups ("Unknown 1" and "Unknown 2"), and chromosome XI was missing (Su *et al.* 2002). A
758 clearer picture emerged when this linkage map was integrated with shotgun sequence assembly
759 data using 1st generation sequencing, leading to a consensus karyotype of 14 chromosomes. In
760 all three of these assemblies (from strains GT1, ME49 and VEG), chromosomes VIIb and VIII
761 always assembled into separate contigs. However, in some studies significant genetic linkage
762 between markers on former chromosomes VIIb and VIII was observed (for example see the
763 discussion in (Khan *et al.* 2005) and figure S2 in (Khan *et al.* 2014)), but the lack of any
764 contiguous assemblies for these two genome fragments (as found on ToxoDB.org as well as
765 outlined in (Lorenzi *et al.* 2016a)) led to continued acceptance of the 14 chromosome model. It
766 was only until recently where reports using chromosome-capture technologies ("Hi-C"; (Bunnik
767 *et al.* 2019)) suggested a fusion between the VIIb and VIII genome fragments, and this was the
768 first study to propose a 13 chromosome karyotype that is most consistent with our nuclear

genome assembly..

Our work here conclusively establishes the *T. gondii* 13 chromosome karyotype in *T. gondii* strain ME49, and show that this karyotype is conserved across all queried *T. gondii* strains and one of its near relatives, *Neospora caninum*. The reason for the consistent prediction that fragments VIIb and VIII were distinct in *T. gondii* was clearly due to repetitive sequences near the breakpoint (hence the consistent and artificial fragmentation of this chromosome across multiple de novo sequenced strains using both first and second generation sequencing technology (Lorenzi *et al.* 2016a)). For *N. caninum*, it appears that the 14 chromosome model was based on overestimation of the degree of synteny between *T. gondii* and *N. caninum* (Reid *et al.* 2012). This karyotype that is robustly supported by our assemblies are consistent with existing Hi-C data (Bunnik *et al.* 2019) and existing genetic linkage maps from F1 progeny derived from type IxII and IxIII crosses (Khan *et al.* 2005; Khan *et al.* 2014).

Tandem gene expansion followed by selection-driven diversification provides a means for genome innovation and neofunctionalization, and this has occurred at multiple loci in the *T. gondii* genome (Adomako-Ankomah *et al.* 2014; Adomako-Ankomah *et al.* 2016; Blank and Boyle 2018). These loci can differ in copy number between strains, including those belonging to the same clonal lineage (for example MAF1 and ROP5 copy number differs between “Type 1” strains GT1 and RH (Adomako-Ankomah *et al.* 2014; Adomako-Ankomah *et al.* 2016), and MAF1 copy number differs between “Type 3” strains CTG and VEG (Adomako-Ankomah *et al.* 2016). While it is generally assumed that copy number changes can occur during errors in DNA replication, this could occur with different frequency during sexual versus asexual propagation. Here we show that some loci can change in gene number and content during sexual recombination by sequencing multiple F1 progeny from a well-defined cross between type 2 and type 3 *T. gondii*. Specific changes in copy number and/or content at specific loci could have a dramatic impact on the overall virulence phenotypes of individual F1 progeny that emerge from natural crosses.

795 Our long-standing interest in tandem gene arrays was one of the original motivations of
796 this study as these loci are particularly prone to incomplete and inaccurate assembly in a variety
797 of species including *T. gondii* and N. caninum (Adomako-Ankomah *et al.* 2014; Jain *et al.* 2018;
798 Lapp *et al.* 2018). While the above analysis was sufficient to accurately determine the number,
799 order and orientation of these genes in multiple strains of *T. gondii* including the F1 progeny
800 clones, the loci were still artifactually pseudogenized even after multiple rounds of polishing with
801 Pilon. We are not aware a comprehensive attempt to solve this problem, possibly because
802 single copy genes tend to be very accurately corrected using tools like Pilon (Senol Cali *et al.*
803 2019). As a case in point, a hybrid PacBioRS/Hi-C assembly of the *P. knowlesi* genome (Lapp
804 *et al.* 2018) required manual assembly correction to removed hundreds of incorrectly
805 pseudogenized members of the *SICAvar* gene family. Our approach here relies on a similar
806 approach but is automated to permit multiple rounds of error correction efficiently using a multi-
807 core computer running Linux. Based on our analyses of these sequences they appear to be the
808 most precise version of these intractable sequences to date, given that current versions of these
809 genes deposited to Genbank from prior publications from our group and others on ROP5
810 (Reese et al. 2011) and MAF1 (Adomako-Ankomah et al. 2014; Adomako-Ankomah et al. 2016)
811 were obtained by PCR and subject to artefactual chimerism. This is particularly evident for the
812 MAF1 locus which is made up of tandem expansions of a 2 gene cassette (which we have
813 dubbed MAF1A and MAF1B and these genes are functionally distinct (only MAF1B drives the
814 unique T. gondii host mitochondrial association phenotype (Adomako-Ankomah et al. 2016;
815 Blank et al. 2018)).
816 Taken together these data clearly demonstrate that the *T. gondii* karyotype has been
817 mischaracterized as being made up of 14, rather than 13, chromosomes. The 13 chromosome
818 karyotype is consistent across multiple strains and is also conserved across the local
819 phylogeny, in that *N. caninum* has the same number. This is consistent with a propensity for
820 chromosome number to be largely conserved between closely-related species even when gene

821 content and overall synteny may differ dramatically (as they do for *N. caninum* and *T. gondii*, in
822 contrast to previous work that overestimated the extent of synteny between these species; (Reid
823 *et al.* 2012)). These data represent a new era in genome sequencing in *T. gondii* and its near
824 relatives, allowing for near-complete telomere-to-telomere assemblies of *T. gondii* strains to be
825 generated with minimal effort and cost. Moreover with existing and supplementary correction
826 methods that are targeted to the systematic error that is most common in Oxford Nanopore
827 sequencing data (incorrect calls of homopolymer nucleotide runs), we have been able to
828 generate the most accurate version of a key subset of virulence effector genes in this organism
829 with such an enormous impact on human global heatlh.
830

831 **METHODS**

832 **Parasite and cell culture**

833 All *T. gondii* strains and *N. caninum* Liverpool strain were maintained by serial passage of
834 tachyzoites in human foreskin fibroblasts (HFFs). HFFs were cultured in Dulbecco's modified
835 Eagle's medium (DMEM) containing 10% fetal bovine serum (FBS), 2 mM glutamine, and 50
836 mg/ml each of penicillin and streptomycin at 37 ℃ in a 5% CO_2 incubator.

837 **High-molecular-weight (HMW) genomic DNA extraction**

838 Prior to DNA purification, tachyzoites of *T. gondii* or *N. caninum* Liverpool strain were grown
839 in 2×10^7 HFFs for about 5-7 days until the monolayer was fully infected. The infected cells were
840 then scraped, and syringe-lysed to release the parasites, and the parasites were harvested by
841 filtering (5.0 μm syringe filter, Millipore) and centrifugation. The pelleted parasites were
842 resuspended and lysed in 10 ml TLB buffer (100 mM NaCl, 10 mM Tris-HCl [pH 8.0], 25 mM
843 EDTA [pH 8.0], 0.5% (w/v) SDS) containing 20 μg/ml RNase A for 1 hour at 37 ℃, followed by a
844 3-hour proteinase K (20 mg/ml) digestion at 50 ℃. The lysate was split into two tubes containing
845 phase-lock gel (Quantabio), and 5 ml TE-saturated phenol (Millipore Sigma) was added to each
846 tube, mixed by rotation for 10 min and centrifuged for 10 min at 4,750×g. The DNA was isolated

847 by removing the aqueous phase to two tubes containing phase-lock gel, followed by a 25:24:1
848 phenol-chloroform-isoamyl alcohol (Millipore Sigma) extraction. The DNA in the aqueous phase
849 was further purified by ethanol precipitation by adding 4 ml 3 M NaOAc (pH 5.2), and then
850 mixing 30 ml ice-cold 100% ethanol. The solution was mixed by gentle inversion and briefly
851 centrifuged at 1000×g for 2 min to pellet the DNA. The resulting pellet was washed three times
852 with 70% ethanol, and all visible traces of ethanol were removed from the tube. The DNA was
853 allowed to air dry for 5 min on a 40 ℃ heat block, resuspended in 40 µl elution buffer (10 mM
854 Tris-HCl [pH 8.5]) without mixing on pipetting, followed by an overnight incubation at 4 ℃. The
855 concentration and purity of the eluted DNA were measured using a NanoDrop
856 spectrophotometer (Thermo Scientific), and approximately 400 ng of DNA was used for
857 sequencing library preparation.
858
859 **MinION library preparation and sequencing**
860 The MinION sequencing libraries were prepared using the SQK-RAD004 or SQK-RBK004
861 kit (Oxford Nanopore Technologies) protocol accompanying all pipetting steps performed using
862 pipette tips with ~1 cm cut off of the end. High molecular weight DNA (7.5 µl corresponding to
863 400ng of DNA) was mixed with 2.5 µl of fragmentation mix (SQK-RAD004 kit) or barcoded
864 fragmentation mix (SQK-RBK004 kit), and then incubated at 30 ℃ for 1 min, followed by 80 ℃
865 for 1 min on a thermocycler. After incubation, 1 µl of rapid adapter mix was added and mixed
866 gently by flicking the tube, and the library was incubated at room temperature for 5 min. Prior to
867 the library loading, the flow cell (MinION R9.4.1 flow cell [FLO-MIN106, Oxford Nanopore
868 Technologies]) was primed by loading 800 µl of priming mix (flush tether and flush buffer mix,
869 Oxford Nanopore Technologies) into the priming port on the flow cell and left for 5 min. After
870 priming, 11 µl of DNA library was mixed with 34 µl of sequencing buffer (Oxford Nanopore
871 Technologies), 25.5 µl of resuspended loading beads (Oxford Nanopore Technologies), and 4.5
872 µl of nuclease-free water. To initiate sequencing, 75 µl of the prepared library was loaded onto

873 the flow cell through the SpotON sample port in a drop-by-drop manner. Sequencing was
874 performed immediately after platform QC, which determined the number of active pores. The
875 sequencing process was controlled using MinKNOW (Oxford Nanopore Technologies) and the
876 resulting FAST5 files were base-called using Guppy v3.2.1 (Oxford Nanopore Technologies).
877 The barcoded sequencing reads were demultiplexed using Deepbinner
878 (https://github.com/rrwick/Deepbinner). Read statistics were computed and graphed using
879 Nanoplot v1.0.0 (De Coster *et al.* 2018).

880 **Read quality control and *de novo* genome assembly**

881 To assess read quality, raw sequencing reads were aligned against the reference genomes
882 (Information of the reference genomes used in this study are shown in **Table S1**) using
883 Minimap2 (Li 2018) with the following parameter: -ax map-ont. All reads > 1000 bp in length
884 were input into Canu v1.7.1 (Koren *et al.* 2017) for *de novo* assembly using the complete Canu
885 pipeline (correction, trimming, and assembly) (Koren *et al.* 2017) with the following parameters:
886 correctedErrorRate=0.154, gnuplotTested=TRUE, minReadLength=1000, and –nanopore-raw.
887 Assembly was performed based on an estimated 65 Mb genome size for *T. gondii* strains, and
888 57 Mb for *N. caninum* Liverpool strain, and was run using the Slurm management system on the
889 high throughput computing (HTC) cluster at University of Pittsburgh.

890 **Error correction and assembly polishing**

891 For the Canu-yielded *Tg*RH88, *Tg*ME49, *Tg*CTG, and *Nc*Liv assemblies, assembly errors
892 were corrected by Pilon v1.23 (Walker *et al.* 2014) with four iterations using the alignment of
893 select whole-genome Illumina paired-end reads (SRA: SRR5123638, SRR2068653,
894 SRR5643140, or ERR701181) to the assembly contigs generated by BWA-MEM (Li 2013). The
895 resulting corrected contigs were reassembled using Flye v2.5 (Kolmogorov *et al.* 2019). For the
896 II×III F1 progeny assemblies, CL13, S27, S21, S26, and D3X1, the assembly contigs were
897 directly subjected to Flye for reassembly without Pilon correction. The final contigs/scaffolds in
898 the *Tg*RH88, *Tg*ME49, and *Tg*CTG assemblies were assigned, ordered, and oriented to

899 chromosomes using ToxoDB-48 genomes as reference (**Table S1**). These genomes were then
900 deposited in Genbank.

901 **Supplemental assembly correction**

902 Since tandem gene arrays were still artificially pseudogenized after Pilon-based error
903 correction we performed multiple rounds of error correction to eliminate any remaining
904 homopolymer errors. First we aligned the Illumina sequence reads used in Pilon to the polished
905 assembly using bowtie2 (using the -k 10 parameter to allow reads to map to up to 10 distinct
906 locations). We then broke the genome into 250 Kb fragments and counted all possible 30 bp
907 kmers in the raw reads aligning to that 250 Kb region. Then, we walked through the assembly 1
908 bp at a time, extracting the 30 bp starting at that position and counting the number of times that
909 30 bp was found in the reads mapping to the 250 Kb region. If the read count was < 5 we
910 attempted to determine the length of the homopolymer by adding sequence iteratively until the
911 number of reads harboring that sequence increased above 10. The nucleotide to be added was
912 selected only of ≥90% of the reads had the same nucleotide at that position. If the read was not
913 improved after addition of 10 nucleotides, no changes were made to the sequence, unless the
914 truncation of 1 nucleotide from the end of the 30 bp assembly fragment increased coverage, in
915 which case that change was made to the assembly. We repeated this correction 4 times
916 iteratively, and used the corrected assemblies to determine the sequence of individual paralogs
917 at tandem gene arrays such as MAF1, ROP5 and ROP38.

918

919 **Long-read assembly evaluation**

920 Assembly statistics were computed using Canu v1.7.1 and QUAST v5.0.2 (Mikheenko *et al.*
921 2018). Genome assembly completeness assessment was performed using BUSCO v3.0.2
922 (Waterhouse *et al.*) against the Protists ensembl dataset. Gene predictions were performed
923 using Augustus v3.3 (Keller *et al.* 2011) with the *Toxoplasma gondii*-specific training set.

924

925 **Whole genome alignment**

926 Whole genome alignments between the long-read assemblies and the reference genomes
927 were performed using MUMmer v4.0.0 (Kurtz *et al.* 2004) and Mauve v2.4.0 (Darling *et al.*
928 2004). Dotplots were generated using D-Genies (Cabanettes and Klopp 2018). BWA-MEM was
929 used for remapping the corrected reads to the reference genomes, and all SAM files were
930 parsed to sorted BAM files using SAMtools v1.9 (Li *et al.* 2009). Alignments were visualized
931 using a variety of tools including IGV v2.4.15 (Thorvaldsdottir *et al.* 2013), Mauve (Darling *et al.*
932 2004), Circos (Krzywinski *et al.* 2009) and custom scripts implemented in R statistical software
933 (Team 2020).

934 **Structural variant detection**

935 Structural variant differences between the long-read assemblies and reference genomes
936 (from ToxoDB.org or Genbank, depending on the strains analyzed) were identified by
937 processing the delta file generated by the MUMmer alignment generator NUCmer with the
938 parameter “show-diff”. In addition, manual curation of structural variants was performed by
939 visual inspection of chromosomal rearrangements based on the whole genome alignments
940 generated using Mauve and MUMmer, and using BLASTN to identify and count repetitive loci.
941 Select alignment plots were generated to integrate these data using R statistical software (Team
942 2020).

943 **Copy number variant detection**

944 For all gene-coding tandem expansions in *T. gondii* or *N. caninum* identified previously
945 (Adomako-Ankomah *et al.* 2014), all predicted gene sequences were downloaded from
946 ToxoDB.org and aligned using BLASTN, and alignments showing > 80% identity and covering
947 at least 80% of the query gene were used to count tandem repeat numbers in *T. gondii* strains
948 RH88, ME49 and CTG. For a subset of these tandemly expanded loci (*ROP5*, *ROP38*, *MIC17*,
949 *MAF1*, *ROP4/7*, and *TSEL8*), similar analyses were performed manually against all queried *T.*
950 *gondii* strains (including F1 progeny clones), and in this case only alignments that showed more

951 than 95% identity and 98% coverage were considered as a match. Paralogs were grouped
952 based on alignment identity, and the number of copies at these loci was estimated by alignment
953 match counts. Only matches that were within a single assembled Nanopore-derived contig were
954 considered for copy number estimates, and the length of the sequence between the upstream of
955 the first match or the downstream of the last match on the genomic coordinate and the edge of
956 the corresponding contig had to be longer than that of the sequence between two adjacent
957 matches.

958 **Identification and analysis of new sequences in the long-read assemblies**

959 To identify new sequences that filled reference assembly gaps, we aligned the long-read
960 chromosome contigs to the reference assembly chromosomes using NUCmer with the "show-
961 diff -q" parameter. The coordinates of a) sequence expansions and b) unaligned sequences
962 from our *de novo* assembly were determined, and the corresponding sequences were extracted
963 using custom scripts. Repeats in all genome assemblies were detected using Tandem Repeat
964 Finder (TRF) v4.09 (Benson 1999) and all repeats with a period size of ≥ 500 bp and having at
965 least 2 copies were used to determine the impact of long read assembly on resolution of repeats
966 > 500 bp in size (which are poorly resolved by 2nd or 3rd generation sequencing technologies.

967 **Hi-C Data analysis**

968 Published Hi-C reads (Bunnik *et al.* 2019) were realigned to assemblies TgRH88,
969 TgME49, TgCTG, S27, and S21, then processed further by assigning fragments and removing
970 invalid and duplicate pairs using the processing pipeline HiCPro (Servant *et al.* 2015). Resulting
971 raw intrachromosomal and interchromosomal contact maps were built at 10-kb resolution and
972 corrected for experimental and technical biases using ICE normalization (Imakaev *et al.* 2012).

973 **DATA ACCESS**

974 All raw sequence reads (as fastq files) and polished assemblies have been deposited at
975 Genbank under BioProject number PRJNA638608.

976 **ACKNOWLEDGEMENTS**

977 The authors would like to thank members of the Boyle lab for critical reading of the
978 manuscript and Josh Quick for public sharing of protocols for DNA isolation that maximize read
979 length. This work was supported by grants R01AI116855 and R01AI114655 to J.P.B., State 554
980 Scholarship Fund from the China Scholar Council (201708440340) to JX, grant R21 AI142506
981 (NIH) and NIFA-Hatch-225935 (University of California, Riverside) to KGLR.
982

983 **DISCLOSURE DECLARATION**

984 The authors have no conflicts of interest.

985 **TABLES**

986 **Table 1. Description of the *T. gondii* and *N. caninum* strains sequenced in this study.**

Species	Strain	Genotype (ToxoDB PCR-RFLP genotype)	Geographical origin	Host	References
Toxoplasma gondii	RH88	Type I (ToxoDB #10, Type I)	USA	Human	Sabin, 1941
Toxoplasma gondii	ME49	Type II (ToxoDB #1, Type II)	USA	Sheep	Kasper and Ware, 1985
Toxoplasma gondii	CTG	Type III (ToxoDB #2, Type III)	USA	Cat	Pfefferkorn et al., 1977
Toxoplasma gondii	CL13	Types II×III F1 progeny	USA	Cat	Sibley *et al.*, 1992
Toxoplasma gondii	S27	Types II×III F1 progeny	USA	Cat	Sibley *et al.*, 1992
Toxoplasma gondii	S21	Types II×III F1 progeny	USA	Cat	Sibley *et al.*, 1992
Toxoplasma gondii	S26	Types II×III F1 progeny	USA	Cat	Sibley *et al.*, 1992
Toxoplasma gondii	D3X1	Types II×III F1 progeny	USA	Cat	Saeij *et al.*, 2007
Neospora caninum	Liverpool	-	USA	Dog	Dubey *et al.*, 1988

987

988

989

990 **Table 2. Metrics of long-read assemblies and the reference genomes after polishing.**

	Reference genome		Long-read assembly in this study								
	ToxoDB-48 *Tg*ME49	**ENA *Nc*Liv**	***Tg*RH88**	***Tg*ME49**	***Tg*CTG**	**F1_CL13**	**F1_S27**	**F1_S21**	**F1_S26**	**F1_D3X1**	***Nc*Liv**
contigs	2511	247	23[1]	38	29	36	59[2]	32	236	39	58
scaffold gaps	267	234	2	0	0	0	1	0	0	0	0
Total bases (bp)	65,464,242	57,524,120	64,918,878	64,923,798	64,731,950	64,701,501	64,164,327	63,627,129	60,431,397	63,972,942	62,076,476
Maximum contig length (bp)	4,347,958	1,378,223	12,055,564	12,088,238	12,092,387	8,806,958	12,022,215	11,984,269	1,624,740	11,993,556	10,862,166
Contig length N50 (bp)	1,219,553	405,161	6,778,623	6,675,137	6,680,781	3,584,337	6,640,858	6,624,864	476,673	5,127,130	6,406,690
L50	17	50	4	4	4	6	4	4	36	4	4
Contig length N75 (bp)	609,575	224,642	3,884,487	3,410,322	3,832,413	2,159,680	2,970,120	3,316,180	264,689	3,250,471	3,004,054
L75	35	97	7	8	7	12	8	8	78	9	9
GC (%)	52.29	54.85	52.46	52.35	52.34	52.44	52.44	52.42	52.31	52.37	54.66
Length(s) of contigs with Plastid genome sequence	35,372	ND	34,855	118,556	214,201	86,161	34,657	33,601	8,902	127,818	35,072

[1]20 contigs and 1 scaffold

[2]57 contigs and 1 scaffold

991

992

Table 3. Metrics of the long-read assemblies before and after polishing.

	*Tg*RH88		*Tg*ME49		*Tg*CTG	
	Initial assembly	**Final assembly**	**Initial assembly**	**Final assembly**	**Initial assembly**	**Final assembly**
Contiguity						
# contigs/scaffolds	23	21	38	38	38	29
Total bases (bp)	64,401,064	64,918,878	64,522,756	64,923,798	64,789,158	64,731,950
Maximum contig length(bp)	11,930,269	12,055,564	12,002,493	12,088,238	12,040,189	12,092,387
Contig/scaffold length N50 (bp)	6,718,904	6,778,623	6,635,075	6,675,137	6,653,560	6,680,781
Accuracy						
Genome fraction (%)	97.514	97.470	95.914	95.915	98.367	98.353
# mismatches per 100 Kbp	85.91	57.33	61.34	49.82	47.85	44.11
# indels per 100 Kbp	706.9	35.1	537.5	26.59	348.43	21.94
Largest alignment (bp)	2,737,008	2,768,614	4,425,732	4,452,657	2,573,606	2,584,520
Total aligned length (bp)	62,839,196	63,538,670	63,828,373	64,235,943	63,473,895	63,611,317
Completeness						
Complete BUSCOs protists (%)	23.3	88.9	39.1	91.6	54.0	91.2
Fragmented BUSCOs protists (%)	0.9	0.0	1.4	0.0	1.9	0.0
Missing BUSCOs protists (%)	75.8	11.1	59.5	8.4	44.1	8.8

993

994

995 **REFERENCES:**

996 Adomako-Ankomah Y, English ED, Danielson JJ, Pernas LF, Parker ML, Boulanger MJ, Dubey JP, Boyle
997 JP. 2016. Host Mitochondrial Association Evolved in the Human Parasite Toxoplasma gondii via
998 Neofunctionalization of a Gene Duplicate. *Genetics* **203**: 283-298.
999 Adomako-Ankomah Y, Wier GM, Borges AL, Wand HE, Boyle JP. 2014. Differential locus expansion
1000 distinguishes Toxoplasmatinae species and closely related strains of Toxoplasma gondii. *mBio* **5**:
1001 e01003-01013.
1002 Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*
1003 **27**: 573-580.
1004 Blank ML, Boyle JP. 2018. Effector variation at tandem gene arrays in tissue-dwelling coccidia: who
1005 needs antigenic variation anyway? *Current opinion in microbiology* **46**: 86-92.
1006 Blank ML, Parker ML, Ramaswamy R, Powell CJ, English ED, Adomako-Ankomah Y, Pernas LF,
1007 Workman SD, Boothroyd JC, Boulanger MJ et al. 2018. A Toxoplasma gondii locus required for
1008 the direct manipulation of host mitochondria has maintained multiple ancestral functions. *Mol*
1009 *Microbiol* **108**: 519-535.
1010 Brooks CF, Francia ME, Gissot M, Croken MM, Kim K, Striepen B. 2011. Toxoplasma gondii sequesters
1011 centromeres to a specific nuclear region throughout the cell cycle. *Proceedings of the National*
1012 *Academy of Sciences of the United States of America* **108**: 3767-3772.
1013 Bunnik EM, Venkat A, Shao J, McGovern KE, Batugedara G, Worth D, Prudhomme J, Lapp SA, Andolina
1014 C, Ross LS et al. 2019. Comparative 3D genome organization in apicomplexan parasites.
1015 *Proceedings of the National Academy of Sciences of the United States of America*
1016 doi:10.1073/pnas.1810815116.
1017 Burg JL, Grover CM, Pouletty P, Boothroyd JC. 1989. Direct and sensitive detection of a pathogenic
1018 protozoan, Toxoplasma gondii, by polymerase chain reaction. *Journal of clinical microbiology* **27**:
1019 1787-1792.
1020 Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple
1021 way. *PeerJ* **6**: e4958.
1022 Clemente M, de Miguel N, Lia VV, Matrajt M, Angel SO. 2004. Structure analysis of two Toxoplasma
1023 gondii and Neospora caninum satellite DNA families and evolution of their common monomeric
1024 sequence. *J Mol Evol* **58**: 557-567.
1025 Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic
1026 sequence with rearrangements. *Genome Res* **14**: 1394-1403.
1027 De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and
1028 processing long-read sequencing data. *Bioinformatics (Oxford, England)* **34**: 2666-2669.
1029 Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence
1030 sets. *Curr Protoc Bioinformatics* **Chapter 10**: Unit 10 13.

1031 Diaz-Viraque F, Pita S, Greif G, Moreira de Souza RdC, Iraola G, Robello C. 2018. Nanopore sequencing
1032 significantly improves genome assembly of the eukaryotic protozoan parasite Trypanosoma cruzi.
1033 *bioRxiv* doi:10.1101/489534: 489534.
1034 Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander
1035 ES, Aiden AP et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields
1036 chromosome-length scaffolds. *Science* **356**: 92-95.
1037 Echeverria PC, Rojas PA, Martin V, Guarnera EA, Pszenny V, Angel SO. 2000. Characterisation of a
1038 novel interspersed Toxoplasma gondii DNA repeat with potential uses for PCR diagnosis and
1039 PCR-RFLP analysis. *FEMS Microbiol Lett* **184**: 23-27.
1040 Edvinsson B, Lappalainen M, Evengard B, Toxoplasmosis ESGf. 2006. Real-time PCR targeting a 529-bp
1041 repeat element for diagnosis of toxoplasmosis. *Clinical microbiology and infection : the official*
1042 *publication of the European Society of Clinical Microbiology and Infectious Diseases* **12**: 131-136.
1043 Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities.
1044 *Genome Res* **8**: 186-194.
1045 Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I.
1046 Accuracy assessment. *Genome Res* **8**: 175-185.
1047 Fournier T, Gounot JS, Freel K, Cruaud C, Lemainque A, Aury JM, Wincker P, Schacherer J, Friedrich A.
1048 2017. High-Quality de Novo Genome Assembly of the Dekkera bruxellensis Yeast Using
1049 Nanopore MinION Sequencing. *G3 (Bethesda, Md)* **7**: 3243-3250.
1050 Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ
1051 et al. 2008. ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic acids research*
1052 **36**: D553-556.
1053 Gissot M, Walker R, Delhaye S, Huot L, Hot D, Tomavo S. 2012. Toxoplasma gondii chromodomain
1054 protein 1 binds to heterochromatin and colocalises with centromeres and telomeres at the nuclear
1055 periphery. *PloS one* **7**: e32671.
1056 Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA.
1057 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature*
1058 *methods* **9**: 999-1003.
1059 Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT et
1060 al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature*
1061 *biotechnology* **36**: 338-345.
1062 Jones JL, Holland GN. 2010. Annual burden of ocular toxoplasmosis in the US. *The American journal of*
1063 *tropical medicine and hygiene* **82**: 464-465.
1064 Joynson DH, Wreghitt TG. 2005. *Toxoplasmosis: a comprehensive clinical guide*. Cambridge University
1065 Press.
1066 Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein
1067 multiple sequence alignments. *Bioinformatics (Oxford, England)* **27**: 757-763.

1068 Khan A, Behnke MS, Dunay IR, White MW, Sibley LD. 2009. Phenotypic and gene expression changes
1069 among clonal type I strains of Toxoplasma gondii. *Eukaryotic cell* **8**: 1828-1836.
1070 Khan A, Shaik JS, Behnke M, Wang Q, Dubey JP, Lorenzi HA, Ajioka JW, Rosenthal BM, Sibley LD.
1071 2014. NextGen sequencing reveals short double crossovers contribute disproportionately to
1072 genetic diversity in Toxoplasma gondii. *BMC Genomics* **15**: 1168.
1073 Khan A, Taylor S, Su C, Mackey AJ, Boyle J, Cole R, Glover D, Tang K, Paulsen IT, Berriman M et al.
1074 2005. Composite genome map and recombination parameters derived from three archetypal
1075 lineages of Toxoplasma gondii. *Nucleic acids research* **33**: 2980-2992.
1076 Kim K, Weiss LM. 2004. Toxoplasma gondii: the model apicomplexan. *International journal for*
1077 *parasitology* **34**: 423-432.
1078 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat
1079 graphs. *Nature biotechnology* **37**: 540-546.
1080 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate
1081 long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-
1082 736.
1083 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos:
1084 an information aesthetic for comparative genomics. *Genome Res* **19**: 1639-1645.
1085 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and
1086 open software for comparing large genomes. *Genome biology* **5**: R12.
1087 Lapp SA, Geraldo JA, Chien JT, Ay F, Pakala SB, Batugedara G, Humphrey J, Ma Hc, De BJ, Le Roch
1088 KG et al. 2018. PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C
1089 correction and manual annotation of the SICAvar gene family. *Parasitology* **145**: 71-84.
1090 Lau YL, Lee WC, Gudimella R, Zhang G, Ching XT, Razali R, Aziz F, Anwar A, Fong MY. 2016.
1091 Deciphering the Draft Genome of Toxoplasma gondii RH Strain. *PloS one* **11**: e0157901.
1092 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1093 *arXiv:13033997v2 [q-bioGN]*.
1094 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**:
1095 3094-3100.
1096 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The
1097 Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078-2079.
1098 Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, Karamycheva S, Pinney
1099 D, Brunk BP, Ajioka JW et al. 2016a. Local admixture of amplified and diversified secreted
1100 pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nature*
1101 *Communications* **7**: 10147.
1102 Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, Karamycheva S, Pinney
1103 D, Brunk BP, Ajioka JW et al. 2016b. Local admixture of amplified and diversified secreted
1104 pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nature*

1105 *Communications* **7**: 10147.
1106 Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM.
1107 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics*
1108 **16**: 327.
1109 Matrajt M, Angel SO, Pszenny V, Guarnera E, Roos DS, Garberi JC. 1999. Arrays of repetitive DNA
1110 elements in the largest chromosomes of Toxoplasma gondii. *Genome* **42**: 265-269.
1111 Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018.
1112 High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat*
1113 *Commun* **9**: 541.
1114 Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly
1115 evaluation with QUAST-LG. *Bioinformatics (Oxford, England)* **34**: i142-i150.
1116 Pfefferkorn ER, Pfefferkorn LC. 1976. Toxoplasma gondii: isolation and preliminary characterization of
1117 temperature-sensitive mutants. *Experimental parasitology* **39**: 365-376.
1118 Quick J. 2018. Ultra-long read sequencing protocol for RAD004 V.3. https://www.protocols.io/view/ultra-
1119 long-read-sequencing-protocol-for-rad004-mrxc57n?version_warning=no.
1120 Reese ML, Zeiner GM, Saeij JP, Boothroyd JC, Boyle JP. 2011. Polymorphic family of injected
1121 pseudokinases is paramount in Toxoplasma virulence. *Proceedings of the National Academy of*
1122 *Sciences of the United States of America* **108**: 9625-9630.
1123 Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Könen-Waisman S, Latham SM, Mourier
1124 T, Norton R, Quail MA et al. 2012. Comparative Genomics of the Apicomplexan Parasites
1125 Toxoplasma gondii and Neospora caninum: Coccidia Differing in Host Range and Transmission
1126 Strategy. *PLOS Pathogens* **8**: e1002567.
1127 Reischl U, Bretagne S, Kruger D, Ernault P, Costa JM. 2003. Comparison of two DNA targets for the
1128 diagnosis of Toxoplasmosis by real-time PCR using fluorescence resonance energy transfer
1129 hybridization probes. *BMC infectious diseases* **3**: 7.
1130 Saeij JP, Boyle JP, Boothroyd JC. 2005. Differences among the three major strains of Toxoplasma gondii
1131 and their specific interactions with the infected host. *Trends Parasitol* **21**: 476-481.
1132 Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings*
1133 *of the National Academy of Sciences of the United States of America* **74**: 5463-5467.
1134 Schmidt MH, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Mass J,
1135 Pfaff C et al. 2017. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore
1136 Sequencing. *The Plant cell* **29**: 2336-2348.
1137 Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. 2019. Nanopore sequencing technology and tools for
1138 genome assembly: computational analysis of the current state, bottlenecks and future directions.
1139 *Briefings in bioinformatics* **20**: 1542-1559.
1140 Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. 2015. HiC-

1141 Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**: 259.
1142 Sibley LD, Ajioka JW. 2008. Population structure of Toxoplasma gondii: clonal expansion driven by
1143 infrequent recombination and selective sweeps. *Annual review of microbiology* **62**: 329-351.
1144 Sibley LD, Boothroyd JC. 1992. Construction of a molecular karyotype for Toxoplasma gondii. *Molecular*
1145 *and biochemical parasitology* **51**: 291-300.
1146 Su C, Howe DK, Dubey JP, Ajioka JW, Sibley LD. 2002. Identification of quantitative trait loci controlling
1147 acute virulence in Toxoplasma gondii. *Proceedings of the National Academy of Sciences of the*
1148 *United States of America* **99**: 10753-10758.
1149 Team RC. 2020. R: A language and environment for statistical computing. R Foundation for Statistical
1150 Computing, Vienna, Austria.
1151 Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance
1152 genomics data visualization and exploration. *Briefings in bioinformatics* **14**: 178-192.
1153 Villard O, Candolfi E, Ferguson DJ, Marcellin L, Kien T. 1997. Loss of oral infectivity of tissue cysts of
1154 Toxoplasma gondii RH strain to outbred Swiss Webster mice. *International journal for*
1155 *parasitology* **27**: 1555-1559.
1156 Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, Wenger AM, Concepcion GT,
1157 Kronenberg ZN, Munson KM et al. 2019. Improved assembly and variant detection of a haploid
1158 human genome using single-molecule, high-fidelity long reads. *Annals of human genetics*
1159 doi:10.1111/ahg.12364.
1160 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J,
1161 Young SK et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection
1162 and Genome Assembly Improvement. *PloS one* **9**: e112963.
1163 Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov
1164 EM. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.
1165