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], Karine Le Roch[4], Ferhat Ay[3,5] and Jon P. Boyle[1*]

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

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

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

[4]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, including short tandem repeats and larger tandem gene expansions. For well-known host-targeting effector loci like rhoptry protein 5 (ROP5) and mitochondrial association factor 1 (MAF1), we were also able to accurately determine the precise gene count, order, sequence and orientation. Finally, when we compared the *T. gondii* and *N. caninum* assemblies we found that while the 13 chromosome karyotype was conserved, we determined that previously unidentified large scale translocation events occurred in *T. gondii* and *N. caninum* since their most recent common ancestry.

INTRODUCTION

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

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

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

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

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

RESULTS

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

The majority of the *T. gondii* isolates collected from North America and Europe belong to three predominant clonal lineages, types I, II, and III (Sibley and Ajioka 2008) and RH strain is a representative strain of the type I lineage (Pfefferkorn and Pfefferkorn 1976). While RH strain shows some unique phenotypes including higher growth rate *in vitro* and inability to form cysts (Villard *et al.* 1997; Khan *et al.* 2009) and is most frequently used in *T. gondii* studies, a complete *de novo* assembly for *Tg*RH genome is lacking. Therefore, our initial efforts were to use long read single molecule sequencing to generate a more complete *Tg*RH88 assembly.

To take advantage of the long-read technology, HMW genomic DNA of *Tg*RH88 strain was extracted using an optimized protocol which was originally designed for extraction of gram-negative bacteria and mammalian cell DNA (Quick 2018). This protocol included a phenol-chloroform-isoamyl alcohol extraction followed by ethanol precipitation, and the use of large

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

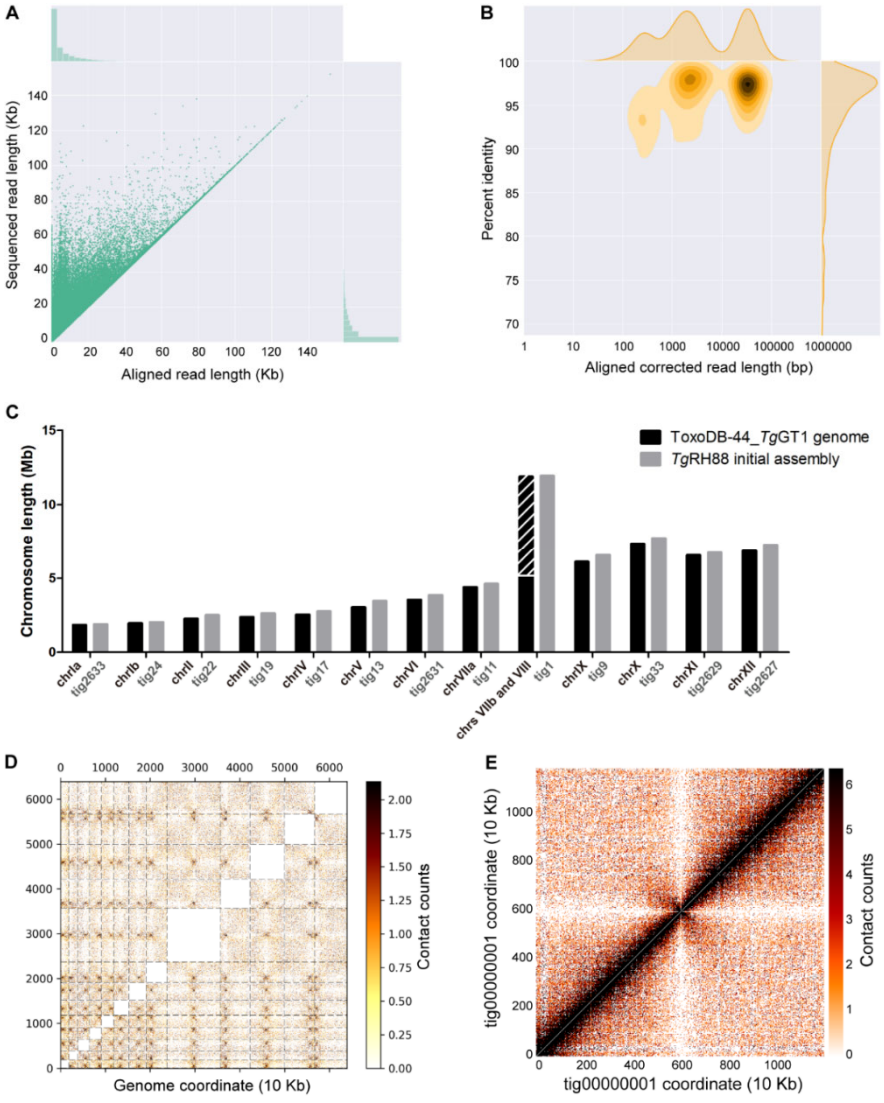


Figure 1. Primary de novo assembly of TgRH88 genome using nanopore reads revises T. gondii karyotype. *(A) Bivariate plot showing a comparison of the aligned read length with the sequencing read length. (B) Bivariate plot*

showing a comparison of the aligned corrected read length (log10 transformed) with the percent identity. (C) Histogram showing comparison of chromosome size between ToxoDB-44_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 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 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 reference genome, we noticed that the sequences annotated as chrVIIb and chrVIII in the ToxoDB-44_*Tg*GT1 genome were parts of a single contig in our *Tg*RH88 assembly (tig1; **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 suggesting 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 TgRH88 tig1 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.

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

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

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

The primary assemblies had a median number of contigs of 38.5 for the *T. gondii* strains and 58 for the *N. caninum* Liverpool strain (**Table S4**). The median assembly size of the *T. gondii* strains was 64 Mb, with a median contig length N50 of 6.63 Mb, and a median L50 of 4 (**Table S4**). The *N. caninum* primary assembly had a size of 61.8 Mb, with a contig N50 size of 6.38 Mb and an L50 of 4 (**Table S4**). Although the mean percent identity of the aligned Canu-corrected reads to the reference genome (~95%) was improved compared to the raw reads (~86%), we used a second round of error correction to improve sequence accuracy. To do this, we mapped whole-genome Illumina paired-end reads (SRA: SRR5123638, SRR2068653, SRR5643140, or ERR701181) to the *Tg*RH88, *Tg*ME49, *Tg*CTG, or *Nc*Liv primary assemblies generated by Canu, and iteratively polished the assembly contigs four times using Pilon v1.23. The polished contigs were then reassembled using Flye v2.5. The resulting contigs/scaffolds were then assigned, ordered, and oriented to chromosomes using relevant ToxoDB-44 reference genomes (**Table S1**). These polished assemblies have been deposited in Genbank (Accession numbers pending).

The final assemblies of *Tg*RH88, *Tg*ME49, or *Tg*CTG consisted of 13 chromosome contigs/scaffolds and varying numbers of unplaced fragments, with an average total size of ~64.8 Mb (**Table 2**). The polished *N. caninum* Liverpool assembly was composed of 58 contigs, showing a cumulative size of 62.1 Mb (**Table 2**). For all species and strains, the long reads and high coverage led to highly improved contiguity of our assemblies compared to the reference genomes. As reported in **Table 2**, with one exception (IIxIII F1 progeny S26), all the *T. gondii* final assemblies were composed of 23-59 contigs, representing a 43-109-fold reduction in the number of contigs in comparison to the ToxoDB-44_*Tg*ME49 assembly (2511 contigs; **Table 2**). For *N. caninum*, compared to the existing assembly of the *Nc*Liv genome (Reid *et al.* 2012), the total number of contigs in the *Nc*Liv final assembly was reduced from 247 to 58 (**Table 2**). The contiguity improvement of long-read assembly was also evident by high contig length N50

values and low L50 values in these assemblies (**Table 2**). Furthermore, for all the long-read assemblies, 55.2% of all chromosomes were fully assembled in single contigs and flanked by two telomeric repeats with proper orientation (e.g., one end 5'-3' TTTAGGG and the other 5'-3' CCCTAAA; **Table S5**). This is in contrast to version 44 of the *T. gondii* ME49 genome housed at ToxoDB where 3 of the 14 chromosome assemblies have telomeric repeats on both ends, 7 have a telomeric repeat on one end, and the remaining 4 putative chromosomes have no telomeric repeats at all. Upon aligning our assemblies to their respective reference genomes (**Table S1**), we found that over 95.9% of the *Tg*RH88, *Tg*ME49, or *Tg*CTG final assembly sequences could be mapped to their relevant reference genome. The alignments revealed that our *de novo* assemblies had fewer mismatches and indels per 100 Kb and higher longest alignment and total aligned length when compared to the assemblies housed on ToxoDB (**Table 3**). To assess genome assembly completeness, we used BUSCO analysis on the polished *Tg*RH88, *Tg*ME49, and *Tg*CTG assemblies and compared the results to a similar analysis for the unpolished assemblies. This analysis, which counts the number of single-copy orthologs unambiguously identified in a genome assembly, found that for 215 such loci 88% of the complete genes were recovered from the polished, long read based assemblies, while 23.3 - 54% were identified in the unpolished assemblies (**Table 3**). These data show the utility of our sequence polishing for improving the accuracy of the assembly.

Table 2. Metrics of long-read assemblies and the reference genomes.

	Reference genome		Long-read assembly in this study								
	ToxoDB-44_*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
[1]20 contigs and 1 scaffold											
[2]57 contigs and 1 scaffold											

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

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

To assess the structural correctness of the long-read assemblies, we aligned our *T. gondii* assembly sequences to the reference genome. As can be seen in **Figure S1**, all of the *T. gondii* long-read assemblies exhibited strong collinearity with their corresponding reference genomes, barring a small number of putative inversions. Consistent with our finding for *Tg*RH88 primary assembly, the chr VIIb/VIII fusion was observed in each of our *T. gondii* assemblies (*Tg*ME49 is represented in **Figure 2A**, red box; and **Figure S1A,B**). To further confirm this observation, we aligned our *Tg*ME49 corrected reads back against the *Tg*ME49 long-read assembly and found an average read depth of 40× for the entirety of tig00000001, and 37× for the "breakpoint" (tig00000001: 5090422 bp) of chrVIIb and chrVIII, indicating that this was unlikely due to an assembly error (**Figure 2B**). We then mapped the corrected nanopore reads again to the ToxoDB-44_*Tg*ME49 reference genome, and the alignments showed that all of the reads that were mapped either to the end of chrVIIb or to the beginning of chrVIII spanned the gap between the two chromosomes, with an average coverage of 105× (**Figure 2C**). This link between reads aligning to chromosome ends was not present in any other chromosome pair (for instance, between chromosomes IX and X; **Figure 2D**).

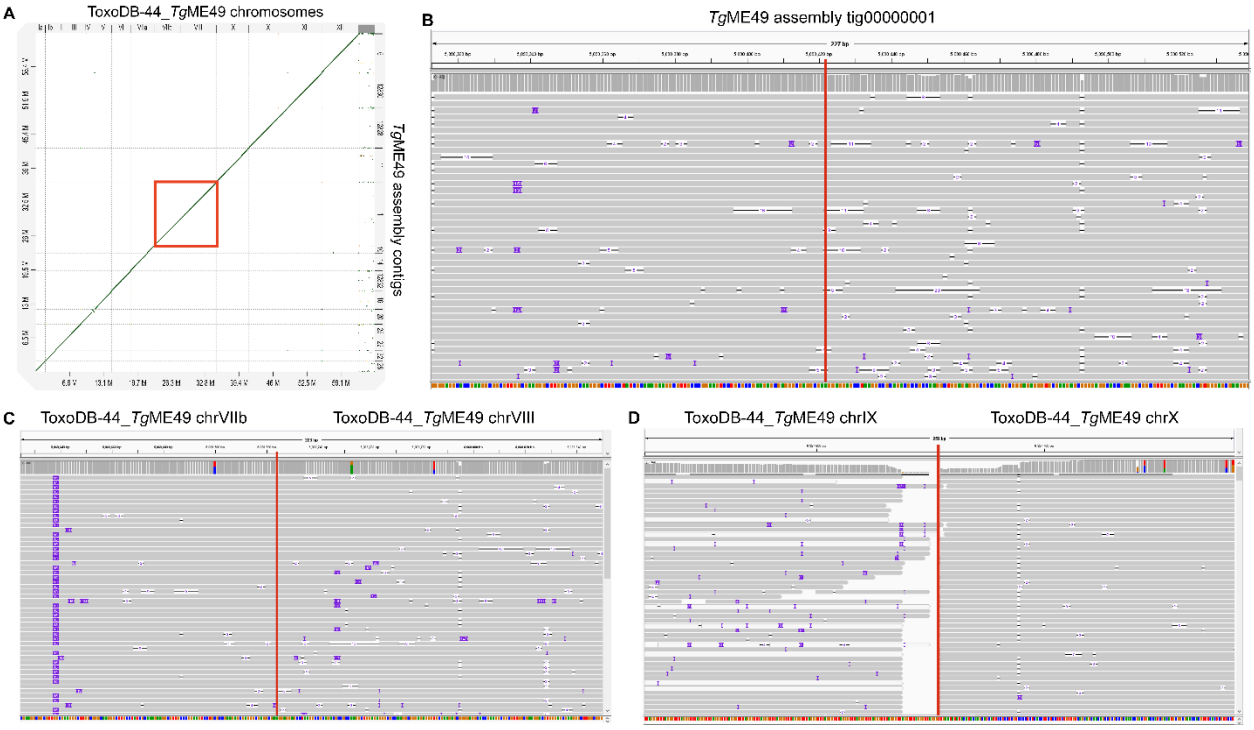


Figure 2. Long-read assembly identifies 13 chromosomes in T. gondii genome. (A) *Dot plot showing the comparison of the TgME49 long-read assembly and the ToxoDB-44_TgME49 genome. Red box shows that the chromosomes VIIb and VIII in the ToxoDB-44_TgME49 genome are fused in a single contig, tig00000001, in the TgME49 long read assembly. (B) Coverage of the "breakpoint" (tig00000001: 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 mapped to the ToxoDB-44_TgME49 genome. (D) Nanopore reads mapping to the end of chromosomes IX and X in the ToxoDB-44-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 tig00000001 was a complete single chromosome (**Figure S2B**). The intrachromosomal contact-count map of tig00000001 showed a

strong and broad diagonal and no aberrant signal along the contig or at the the “breakpoint” of chrVIIb and chrVIII (**Figure S2B**). Similar patterns were also observed in our *Tg*CTG, S27, and S21 assemblies (**Figure S2**). Collectively, these data show the *T. gondii* karyotype has been incorrectly calculated and contains 13, rather than 14, chromosomes. We refer to this fused chromosome as chrVIII in our assembly and have eliminated chrVIIb.

The *Tg*CTG and *Tg*RH88 assemblies had chromosome scale resolution, where 13 contiguous sequences (contigs) corresponded to the 13 chromosomes. However, the S21, S27, and *Tg*ME49 assemblies had several chromosomes composed of two to three large contigs. Hi-C data has historically been used to improve genome assemblies on the basis of contact frequency depending strongly on one-dimensional distance (Dudchenko et al. 2017). That is, Hi-C alignment to contigs in the correct order and orientation would reveal the canonical intrachromosomal pattern of enriched interactions along the diagonal (where one-dimensional genomic distance between bins is the smallest). Hi-C alignment to contigs in the incorrect order and orientation would be illustrated by patterns associated with large-scale inversions, where there is a high interaction frequency between bins placed far away from each other. Leveraging this, we were able to assemble the contigs for chromosomes ChrII, ChrVIIa, ChrIX in S21, ChrIII, ChrIV, ChrV and ChrX in S27, and ChrVIIa in *Tg*ME49. All concatenations have 100 'N' bases in between contigs and are listed below in the order in which they have been concatenated.

For S21 chrII, we inverted tig00000072 and concatenated tig00000072 and tig00000055. For S21 chrVIIa, we concatenated tig00000090 and tig00012823. For S21 chrIX, we concatenated tig00000042 and tig00000018. For S27 chrIII, we inverted tig00000090 and concatenated tig00017745 and tig00000090. For S27 chrIV, we concatenated tig00000192, tig00000094, and tig00000043. For S27 chrV, we concatenated tig00000162 and tig00017744. For S27 chrX, we concatenated tig00000154 and tig00000012. Finally, for *Tg*ME49 chrVIIa, we inverted tig00000014 and concatenated tig00000014 and tig00000010.

309

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

311 Many scaffold gaps within chromosomes remain unresolved (**Table 2**) in the *T. gondii*
312 reference genome particularly in the regions that contain repeated elements. Our long read
313 assembly closed nearly all of the gaps in the *T. gondii* and *N. caninum* genomes (Tables 2 and
314 S6). Furthermore, many unplaced sequences in the *T. gondii* reference genome were
315 assembled into contigs in our assembly. For instance, the unplaced contig KE140372 in the
316 ToxoDB-44_*Tg*ME49 genome, which was 2194 bp in length and contained a sequence
317 encoding a rhoptry protein 4 paralog, was assembled in tig00000028_chrIa in our *Tg*ME49
318 assembly. Unplaced contigs in the ToxoDB-44_*Tg*GT1 genome, AAQM03000823 and
319 AAQM03000824, were assembled in contig_13_chrIII in our *Tg*RH88 assembly.
320 The gap closure enabled us to determine the exact number, order, orientation, and
321 sequence of *T. gondii* duplicated loci like *ROP7*, *ROP5*, *ROP38*, *MIC17*, *MAF1*, and *TSEL8*.
322 The *ROP7* locus is represented in **Figure 3A**, where we identified two unresolved scaffold gaps
323 on chrIa in the ToxoDB-44_*Tg*ME49 genome (black bars), and these gaps marked the site of
324 *ROP4/7* locus. This entire region was spanned by a single contig, tig00000028, in our *Tg*ME49
325 assembly (the blue bar in **Figure 3A**). Aligning the *ROP7* genomic sequence (ToxoDB:
326 TGME49_295110) against tig00000028 using BLASTN revealed 3 copies of the *ROP7* repeat,
327 while only one was predicted in the ToxoDB-44_*Tg*ME49 genome (Figures 3A and 3B)
328 Interestingly, while the ToxoDB-44_*Tg*ME49 genome identified one copy of the *ROP4* gene
329 (GenBank: EU047558.1), our *Tg*ME49 assembly showed that 2 copies of *ROP4* exist this locus,
330 one of which was found between the first and the second copy of *ROP7* (**Figure 3B**). To
331 validate this finding, we identified 12 individual Canu-corrected reads that spanned this entire
332 tandem array, and each one that we examined provided evidence for 3 copies of *ROP7* and 2
333 copies of *ROP4* arranged in the order ROP4_1-ROP7_1-ROP4_2-ROP7_2-ROP7_3 (**Table**
334 **S6**). The copy number and copy order of other known tandem locus expansions (taken from the

supplementary table in (Adomako-Ankomah *et al.* 2014)) in the strains we sequenced were identified and are listed in **Table S6**. Surprisingly, we observed changes in copy number at some of these loci when we compared the parental (*Tg*ME49 or *Tg*CTG) and the progeny (CL13, S27, S21, S26, and D3X1) assemblies (**Figure 3C**). For example, while the *ROP5* locus had 9 copies in our *Tg*ME49 assembly and 4 in our *Tg*CTG assembly, it harbored 6 copies in the F1 progeny S27 and S21, and 7 in S26 (**Figure 3C**). There was an array of 7 tandem copies of *MIC17* in *Tg*ME49 and *Tg*CTG assemblies, whereas it was present in 6 copies in the S26 assembly (**Figure 3C**). These data indicated that changes in copy number and order at tandem gene arrays can occur during sexual recombination. All the new members of the tandem gene arrays and the novel repeated sequences identified here were the result of gap resolution and improved overall contiguity of the genome.

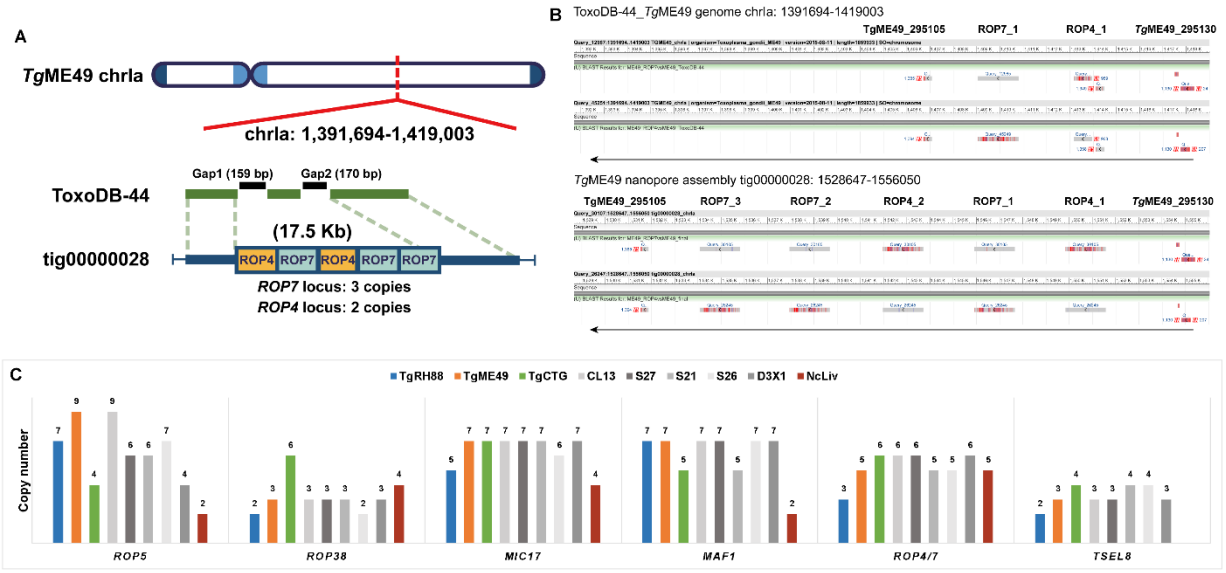


Figure 3. Long-read assembly resolves duplicated locus structure in T. gondii genome. *(A) Two unresolved scaffold gaps on chrIa in ToxoDB-44_TgME49 genome span a 17.5-Kb tandem repeat containing multiple copies of ROP4 and ROP7. The ROP4/7 gaps are closed by the TgME49 long-read assembly (tig00000028), 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-44_TgME49 genome (upper panel) and the TgME49 long-read assembly (lower panel). (C) Precise 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.

Long-read assembly detects structural rearrangements in the *T. gondii* genome

In addition to repeated elements, our assembly was capable of detecting gross structural rearrangements (insertions, deletions, inversions, relocations, and translocations) since the assembly based on long reads was highly contiguous and had near chromosome scale resolution. We aligned the final assemblies of the 8 *T. gondii* strains to the reference genomes and searched for structural variants using eithr Minimap2 (**Figure S1A**) or MUMmer (**Figure S1B**). Consistent with the data shown in **Figure S1**, most contigs in the long-read assemblies were collinear with the chromosomes in the ToxoDB-44 genomes, but not always in a 1:1 correspondence. Interestingly, we observed a 15.7 Kb inversion on chrIII in our *Tg*RH88 assembly, which was absent in *Tg*ME49, *Tg*CTG, or any F1 progeny assembly (**Figure 4A**). We also detected another ~20 Kb inversion on chrXII, which was present in *Tg*ME49, *Tg*CTG, and the F1 progeny assembly, but not in the *Tg*RH88 assembly (*Tg*ME49 is represented in **Figure 4B**).

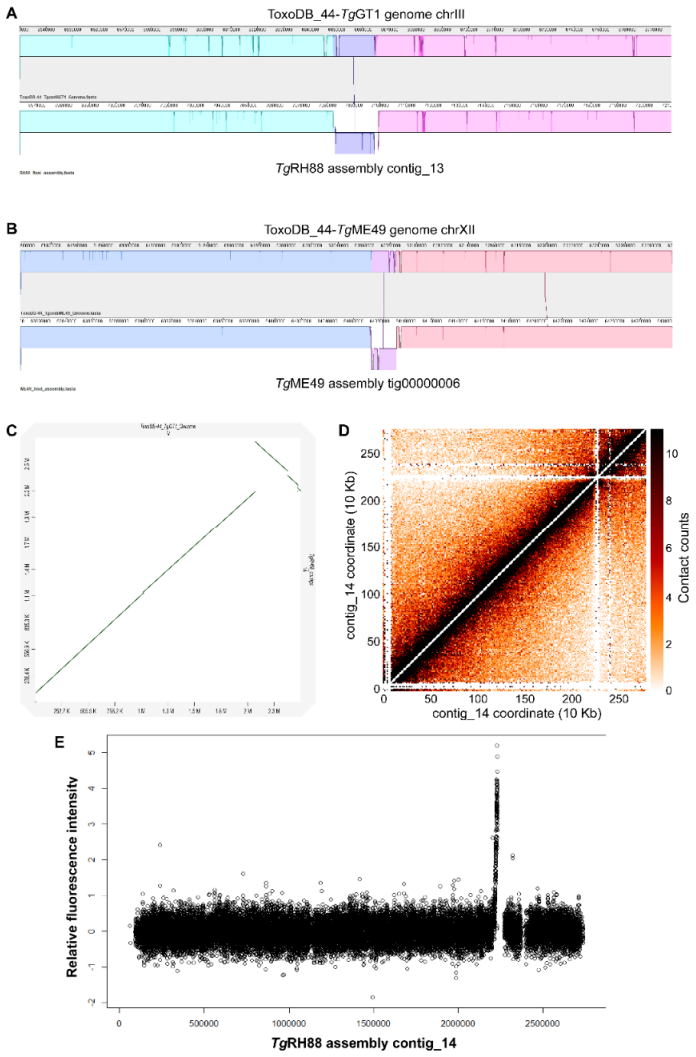


Figure 4. Long-read assembly reveals previously unknown inversions and the centromere location on chrIV in T. gondii. *(A) Inversion in the RH88 long-read assembly on chromosome III relative to the ToxoDB-44-TgGT1 assembly. (B) Inversion in the ME49 long-read assembly on chromosome XII relative to the ToxoDB_44-TgME49 genome. (C) Dot plot comparison of the TgRH88 long-read assembly and the ToxoDB-44_TgGT1 genome showing a 429.3-Kb inversion at 2,096,529-2,525,795 bp on chrIV. (D) Intrachromosomal Hi-C contact-count heat map plotted using the sequence of contig_14 in TgRH88 long-read assembly showing a clear centromere signal at position 2.2-2.3 Mb. (E) ChIP-on-chip signal of centromeric histone 3 variant (CenH3) (Brooks et al. 2011) plotted using the TgRH88 long-read assembly as coordinate.*

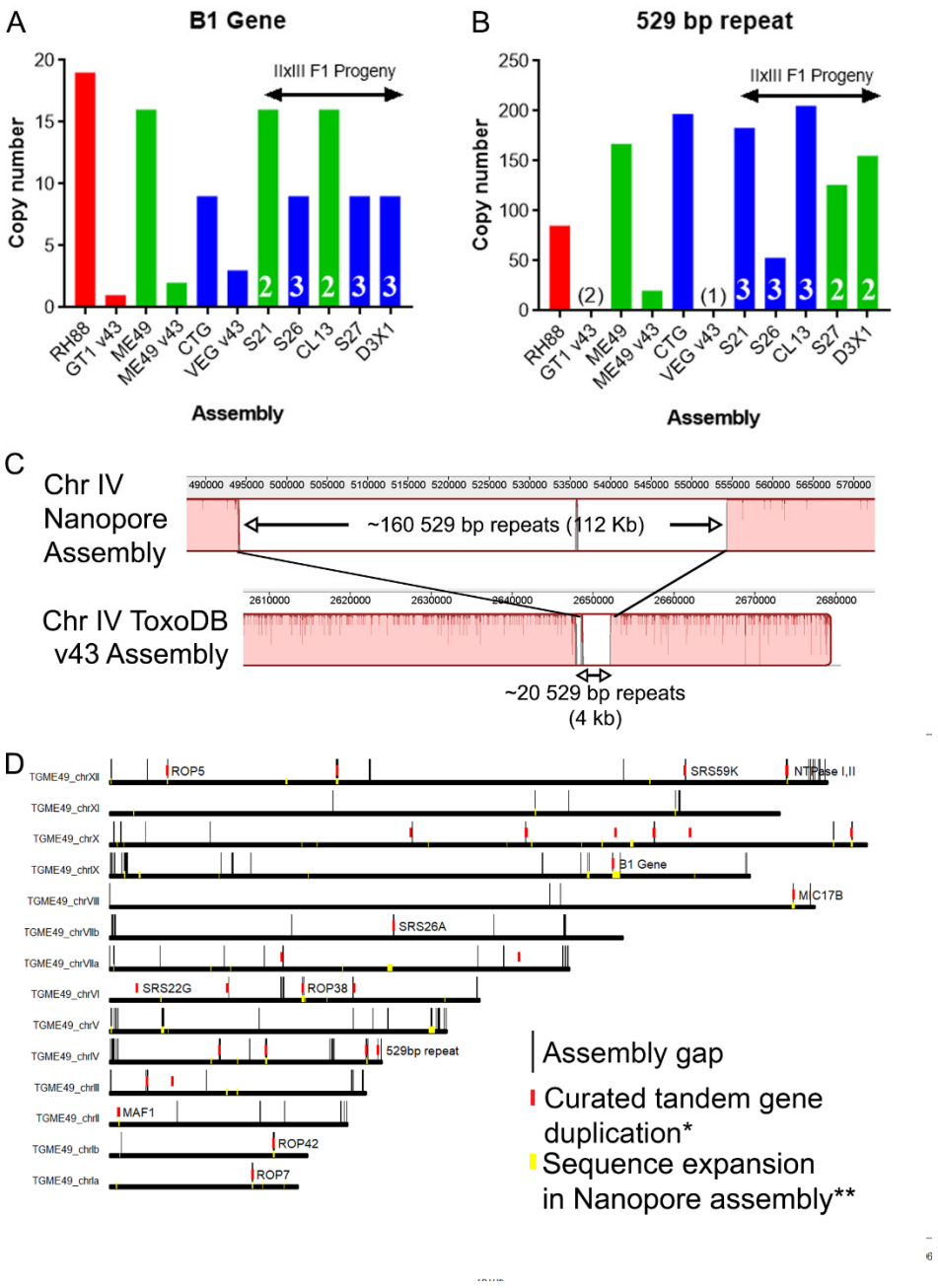
Centromeres for 12 of the 13 *T. gondii* chromosomes have been identified using ChIP-on-chip (chromatin immunoprecipitation coupled with DNA microarrays) of centromeric and peri-centromeric proteins, locations of centromere of chrVIIb and chrIV remain unknown (Brooks *et*

389 *al.* 2011; Gissot *et al.* 2012). For chrVIIb, no hybridization of the centromeric probe to the
390 genomic chip was detected in the ChIP-on-chip assay (Brooks *et al.* 2011), which could be
391 explained by our observation that chrVIIb and chrVIII are a single chromosome, and the
392 centromere of this large chromosome appears to be in the center of this "fused" chromosome, in
393 an area that was previously thought to be the beginning of chromosome VIII (**Figure 1E**). For
394 chromosome IV, two inconsecutive peaks of hybridization were detected at positions 2,501,171-
395 2,527,417 bp on chrIV and 1-9968 bp in the unplaced contig AAQM03000753 in the *Tg*GT1
396 genome based on published ChIP-on-chip data (Brooks *et al.* 2011). Our *Tg*RH88 assembly
397 successfully relocated the sequences in AAQM03000753 into chrIV and revealed a 430.9-Kb
398 inversion event at 2,096,529-2,527,423 bp on chrIV relative to the ToxoDB GT1 reference
399 genome (**Figure 4B**). This inversion was unlikely to be due to an assembly error since it was
400 shown in all of our *T. gondii* assemblies, and was supported by alignment of 128 canu-corrected
401 reads spanning the boundaries of the *Tg*RH88 assembly. When we mapped the ChIP-on-chip
402 data obtained from (Brooks *et al.* 2011) to our *Tg*RH88 assembly (after remapping the probe
403 sequences to our TgRH88 assembly), we resolved the chrIV centromere to a single significant
404 signal peak at 2.20-2.23 Mb on chrIV (**Figure 4C** and **Figure S3**). This finding was also
405 supported by published Hi-C data re-aligned to our Nanopore assemblies, since the
406 intrachromosomal contact count map showed a clear interchromosomal contact signal at 2.20-
407 2.30 Mb on chrIV in the *Tg*RH88 assembly (**Figure 4B**). Collectively, our data not only resolved
408 the molecular karyotype of *T. gondii*, but also resolved the precise location of the chrIV
409 centromere.
410
411 **Long-read assembly adds new sequences to the *T. gondii* reference genome**
412 As shown in **Figure 1C** and **Table S5**, each chromosome-sized contig of our long-read *de*
413 *novo* assemblies was longer than its cognate chromosome of the ToxoDB-44 *T. gondii*
414 reference genome, and for TgRH88, TgME49 and TgCTG we were able to assemble between

1.7 and 3.9 Mb of previously unlocated and/or unassembled sequence to the chromosomes of these assemblies. These new sequences were scattered across the genome, and filled in nearly all of the sequence gaps found in the reference genome. The new sequences added by the long-read assemblies extended the subtelomeric regions of the chromosomes in the *T. gondii* reference genome. While 4 chromosomes of the ToxoDB-44_*Tg*ME49 genome contained no telomeric repeat and 7 were missing one of the telomeric repeats, all of the chromosome contigs in the *Tg*ME49 long-read assembly were assembled up until both telomeric caps (**Table S5**). Both telomeres were found in 12 out of the 13 chromosomes in the *Tg*CTG long-read assembly, and one chromosome contig lacked one of the telomeric repeats, whereas only 5 chromosomes in the ToxoDB-44_*Tg*VEG genome contained one telomere and no telomeric repeat was found in the rest of the chromosomes (**Table S5**). Similarly, both telomeres in 7 chromosomes and one telomere in 6 chromosomes were resolved in the *Tg*RH88 long-read assembly, while there were only 2 chromosomes in ToxoDB-44_*Tg*GT1 genome that contained one telomere (**Table S5**).

In addition to telomeres, the bulk of the remaining new sequence was due to multicopy loci. For example, two repetitive gene sequences are used for high sensitivity detection of *T. gondii* in tissue and environmental samples, the B1 gene (Burg *et al.* 1989) and the so-called “529 bp repeat” (Reischl *et al.* 2003; Edvinsson *et al.* 2006). The precise copy number for these genes has been impossible to determine using 1st and 2nd generation sequencing technologies and/or molecular biological experiments like Southern Blotting. Therefore, we used a curated blastn approach to quantify copy number for each of these sequences across our respective Nanopore assemblies. As shown in figure 5A, copy number for the B1 gene was dramatically higher in our Nanopore assemblies compared to existing ToxoDB assemblies (as expected). What was unexpected, however, was that copy number for this gene was lower than that predicted in the literature, ranging between 9 and 19 tandem copies depending on the strain (**Figure 5A**), compared to quantitative blotting-based estimates of 35 (Burg *et al.* 1989). Copy number at this

441 locus was stable, in that all of the queried IIxIII F1 progeny, copy number for each was identical
442 to the parent from which it obtained that chromosomal segment. In contrast to the B1 locus, the
443 “529 bp repeat” locus varied dramatically between isolates and these same F1 progeny. Copy
444 number ranged from 85 to 205, and copy number at this locus for all F1 progeny varied
445 independently of the underlying genotype for that region (white letters and green/blue in **Figure**
446 **5B**). The size of this genome expansion is best illustrated by the whole chromosome alignment
447 shown in **Figure 5C** comparing the *T. gondii* ME49 529 bp repeat locus in the version 43
448 assembly on ToxoDB to our Nanopore assembly (**Figure 5C**). Importantly the 529 bp repeat
449 locus occurs near a sequence assembly gap, and our long read assembly closed this gap (see
450 below and **Figure 5D**), giving the most accurate estimate of 529 bp repeat copy number in any
451 *T. gondii* strain which again varies compared to estimates in the literature (ranging from 200-300
452 copies; e.g., (Reischl *et al.* 2003; Edvinsson *et al.* 2006)). Regardless, similar to tandem gene
453 arrays discussed above in **Figure 3**, it appears that even noncoding repeats like the B1 gene
454 and the 529 bp repeat can also differ in their capacity to change in number during sexual
455 recombination or not. Moreover, unappreciated strain differences in copy number at these loci
456 may adversely affect interpretation of PCR-based detection assays, especially those using more
457 quantitative methods.



458

459 *Figure 5: Long read sequence assemblies precisely resolve canonical repeat*
460 *sequences and identify additional expansions at gene-harboring loci. (A,B)*
461 *Estimated copy number for Nanopore assemblies and existing genome*
462 *assemblies on ToxoDB ("v43") for T. gondii strain types 1, 2 and 3 and IIxIII*
463 *F1 progeny. In all cases, Nanopore assemblies identified dramatically higher*
464 *numbers of either the B1 gene (A) or the 529 bp repeat (B). In the F1 progeny,*
465 *B1 gene copy number tracked directly with the genotype (type 2 or 3) at that*
466 *locus (A), while these same F1 progeny harbored unique numbers of 529 bp*
467 *copies, all of which were not only distinct from their respective genotypes of*
468 *origin but distinct from one another (B). (C) Whole chromosome alignment*

focused on the 529 bp repeat region for the v43 ToxoDB assembly (bottom) and our Nanopore-based assembly (top). Expansion of the known genome sequence at this locus in the Nanopore sequence compared to the ToxoDB assembly is clear, and consistent with our identification of ~140 previously unknown 529 bp repeats in the ME49 genome. (D) Genome-wide assessment of expanded genome sequences mapped onto version 43 of the T. gondii genome taken from ToxoDB. Black bars indicate sequence gaps, red bars indicate known tandem gene arrays, and yellow bars indicate regions that were expanded by at least 5000 bp in the long read Nanopore assembly.

As described above all of our sequence assemblies increased the size of the contiguous assemblies by 1-3 Mb. While some of this new sequence is most certainly derived from gene-poor regions containing simple tandem repeats, the relatively high gene density of the *T. gondii* genome led us to hypothesize that some of this "new" sequence should be derived from gene sequences that were previously masked by assembly gaps. Therefore we used BLASTN and custom parsing scripts to identify genome expansions in our Nanopore assembly relative to the v43 sequence on ToxoDB specifically using all available predicted genes as query sequences. Overall, we identified 62 gene-containing loci that were at least 10 Kb larger in our Nanopore assembly compared to ToxoDB v43, representing 1.2 Mb of sequence. These expansions are represented in **Figure 5D** as yellow blocks, and are shown along with known tandem gene arrays (red blocks) and existing sequence gaps (black lines). Well known tandem gene arrays that are collapsed in 1st and 2nd generation sequence-based assemblies like MAF1 and ROP5 were identified in this analysis (**Figure 5D**), confirming the accuracy of the approach. What was unexpected was the unequal distribution of these "expansions" in our genome-wide analysis (e.g., compare chromosomes XI and X). These expansions are due to increased copy number as well as previously unknown insertions of repetitive sequence.

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-

44_*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 6A**), 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 6B**). For instance, the tig0000052 in *Nc*Liv long-read assembly contained a portion of chrIX and a portion of chrX in the current *N. caninum* genome (**Figure 6B**). A portion of tig00000006 in *Nc*Liv long-read assembly was mapped to the current *N. caninum* chrX, while the remainder of tig00000006 was mapped to chrIX and a region of it was inverted (**Figure 5B**). 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 6B**). Similar chromosomal rearrangement patterns were observed in other *N. caninum* strains (as described in the co-submitted paper).

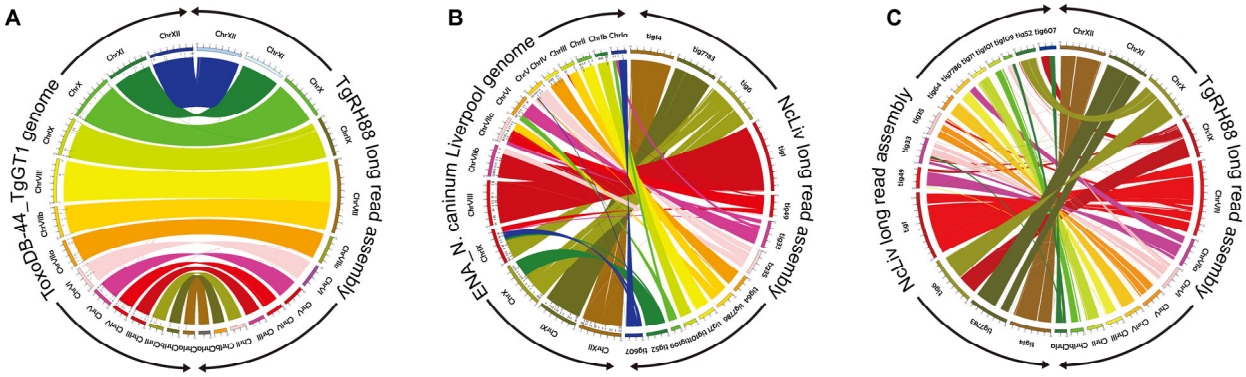


Figure 6. Long-read assembly reveals N. caninum karyotype its synteny with T. gondii. (A) Circos plot showing high synteny between the TgRH88 long-read assembly and the ToxoDB-44_TgGT1 genome. (B) Circos plot showing the chromosomal translocations and inversions in NcLiv long-read assembly compared to the ENA_NcLiv genome. (C) Circos plot showing the syntenic relationship between TgRH88 and NcLiv long-read assembly.

While previous studies (Reid *et al.* 2012; Lorenzi *et al.* 2016) showed that the current *T. gondii* and *N. caninum* reference genomes were highly syntenic, the comparison of *Tg*RH88 long-read assembly with the *Nc*Liv long-read assembly showed smaller blocks of synteny in

most regions between the two genomes (**Figure 5C**). Many individual contigs of the *Nc*Liv long-read assembly were shown to be mapped to multiple chromosome contigs in the *Tg*RH88 long-read assembly. For example, a 3.26-Mb region and a 4.43-Mb region of tig00000006 of the *Nc*Liv long-read assembly were in synteny with regions on contig_1 and contig_3 of *Tg*RH88 long-read assembly, respectively (**Figure 6C**). Moreover, some regions of *Nc*Liv long-read assembly (e.g. tig00007783: 2,000,000-2,500,000 bp) exhibit no synteny with the chromosomes of *Tg*RH88 genome (**Figure 6C**). Overall, our long-read assembly revealed a new and accurate *N. caninum* karyotype, and revised the genomic synteny between the two closely-related species, *T. gondii* and *N. caninum*.

DISCUSSION

The first wave of genome sequencing using first- and second-generation sequencing technologies revolutionized our ability to link phenotype to genotype in diverse strains of *Toxoplasma gondii* and its near relatives. Sequence from multiple isolates have been made publicly available and hosted on ToxoDB.org and were outlined in a recent publication (Lorenzi *et al.* 2016). These genomes, while of great use, were expectedly incomplete due to the presence of hundreds to thousands of sequence assembly gaps depending on sequencing depth and methodology. Recent years have experienced the impact of so-called “3rd generation” technologies that have revolutionized the speed, cost and efficacy of de novo sequence assembly, and most importantly provide a means to dramatically improve existing sequence assemblies. The critical feature of these approaches is the fact that they generate sequence read lengths hundreds to thousands of times longer than those generated using 1st and 2nd generation approaches, and can span repetitive sequence (both short tandem repeats and larger segmental duplications). This technology provides a unique opportunity to greatly improve the assembly of the *T. gondii* genome, particularly at repetitive loci which are known to encode diversified secreted effectors (Adomako-Ankomah *et al.* 2014; Adomako-Ankomah *et al.*

2016).

Our work revises the *T. gondii* and *N. caninum* karyotypes by identifying a previously unappreciated fusion between segments previously thought to be distinct chromosomes (VIIb and VIII). Early genetic mapping experiments and high molecular weight Southern blotting suggested that the *T. gondii* genome was made up of 14 chromosomes (Sibley and Boothroyd 1992; Khan *et al.* 2005). Sequence assemblies provided support for this karyotype, in that genome segments containing genetic markers found on chromosomes VIIb and VIII always assembled into discrete units. Importantly, significant genetic linkage between markers on former chromosomes VIIb and VIII was observed (for example see the discussion in (Khan *et al.* 2005) and figure S2 in (Khan *et al.* 2014)), but the lack of any contiguous assemblies for these two genome fragments (as found on ToxoDB.org as well as outlined in (Lorenzi *et al.* 2016)) led to continued acceptance of the 14 chromosome model. Recent reports using chromosome-capture technologies ("Hi-C"; (Bunnik *et al.* 2019)) also suggested a fusion between the VIIb and VIII genome fragments, and was the first study to propose a 13 chromosome karyotype.

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; (Lorenzi *et al.* 2016)). 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.

Taken together these data clearly demonstrate that the *T. gondii* karyotype has been mischaracterized as being made up of 14, rather than 13, chromosomes. The 13 chromosome karyotype is consistent across multiple strains and is also conserved across the local phylogeny, in that *N. caninum* has the same number. This is consistent with a propensity for chromosome number to be largely conserved between closely-related species even when gene content and overall synteny may differ dramatically (as they do for *N. caninum* and *T. gondii*, in contrast to previous work that overestimated the extent of synteny between these species; (Reid *et al.* 2012)). These data represent a new era in genome sequencing in *T. gondii* and its near relatives, allowing for near-complete telomere-to-telomere assemblies of *T. gondii* strains to be generated with minimal effort and cost.

METHODS

Parasite and cell culture

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

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

603 Prior to DNA purification, tachyzoites of *T. gondii* or *N. caninum* Liverpool strain were grown
604 in 2×10^7 HFFs for about 5-7 days until the monolayer was fully infected. The infected cells were
605 then scraped, and syringe-lysed to release the parasites, and the parasites were harvested by
606 filtering (5.0 μm syringe filter, Millipore) and centrifugation. The pelleted parasites were
607 resuspended and lysed in 10 ml TLB buffer (100 mM NaCl, 10 mM Tris-HCl [pH 8.0], 25 mM
608 EDTA [pH 8.0], 0.5% (w/v) SDS) containing 20 μg/ml RNase A for 1 hour at 37 ℃, followed by a
609 3-hour proteinase K (20 mg/ml) digestion at 50 ℃. The lysate was split into two tubes containing
610 phase-lock gel (Quantabio), and 5 ml TE-saturated phenol (Millipore Sigma) was added to each
611 tube, mixed by rotation for 10 min and centrifuged for 10 min at 4,750×g. The DNA was isolated
612 by removing the aqueous phase to two tubes containing phase-lock gel, followed by a 25:24:1
613 phenol-chloroform-isoamyl alcohol (Millipore Sigma) extraction. The DNA in the aqueous phase
614 was further purified by ethanol precipitation by adding 4 ml 3 M NaOAc (pH 5.2), and then
615 mixing 30 ml ice-cold 100% ethanol. The solution was mixed by gentle inversion and briefly
616 centrifuged at 1000×g for 2 min to pellet the DNA. The resulting pellet was washed three times
617 with 70% ethanol, and all visible traces of ethanol were removed from the tube. The DNA was
618 allowed to air dry for 5 min on a 40 ℃ heat block, resuspended in 40 μl elution buffer (10 mM
619 Tris-HCl [pH 8.5]) without mixing on pipetting, followed by an overnight incubation at 4 ℃. The
620 concentration and purity of the eluted DNA were measured using a NanoDrop
621 spectrophotometer (Thermo Scientific), and approximately 400 ng of DNA was used for
622 sequencing library preparation.
623

MinION library preparation and sequencing

The MinION sequencing libraries were prepared using the SQK-RAD004 or SQK-RBK004 kit (Oxford Nanopore Technologies) protocol accompanying all pipetting steps performed using pipette tips with ~1 cm cut off of the end. Seven point five µl of HMW genomic DNA (corresponding to 400ng of DNA) was mixed with 2.5 µl of fragmentation mix (SQK-RAD004 kit) or barcoded fragmentation mix (SQK-RBK004 kit), and then incubated at 30 ℃ for 1 min, followed by 80 ℃ for 1 min on a thermocycler. After incubation, 1 µl of rapid adapter mix was added and mixed gently by flicking the tube, and the library was allowed to be incubated at room temperature for 5 min. Prior to the library loading, the flow cell (MinION R9.4.1 flow cell [FLO-MIN106, Oxford Nanopore Technologies]) was primed by loading 800 µl of priming mix (flush tether and flush buffer mix, Oxford Nanopore Technologies) into the priming port on the flow cell and left for 5 min. After priming, 11 µl of DNA library was mixed with 34 µl of sequencing buffer (Oxford Nanopore Technologies), 25.5 µl of resuspended loading beads (Oxford Nanopore Technologies), and 4.5 µl of nuclease-free water. To initiate sequencing, 75 µl of the prepared library was loaded onto the flow cell through the SpotON sample port in a drop-by-drop manner. Sequencing was performed immediately after platform QC, which determined the number of active pores. The sequencing process was controlled using MinKNOW (Oxford Nanopore Technologies) and the resulting FAST5 files were base-called using Guppy v3.2.1 (Oxford Nanopore Technologies). The barcoded sequencing reads were demultiplexed using Deepbinner (https://github.com/rrwick/Deepbinner). Read statistics were computed and graphed using Nanoplot v1.0.0 (De Coster *et al.* 2018).

Read quality control and *de novo* genome assembly

To assess read quality, raw sequencing reads were aligned against the reference genomes (Information of the reference genomes used in this study are shown in **Table S1**) using Minimap2 (Li 2018) with the following parameter: -ax map-ont. All reads > 1000 bp in length were input into Canu v1.7.1 (Koren *et al.* 2017) for *de novo* assembly using the complete Canu

pipeline (correction, trimming, and assembly) (Koren *et al.* 2017) with the following parameters: correctedErrorRate=0.154, gnuplotTested=TRUE, minReadLength=1000, and –nanopore-raw. Assembly was performed based on an estimated 65 Mb genome size for *T. gondii* strains, and 57 Mb for *N. caninum* Liverpool strain, and was run using the Slurm management system on the high throughput computing (HTC) cluster at University of Pittsburgh.

Error correction and assembly polishing

For the Canu-yielded *Tg*RH88, *Tg*ME49, *Tg*CTG, and *Nc*Liv assemblies, assembly errors were corrected by Pilon v1.23 (Walker *et al.* 2014) with four iterations using the alignment of select whole-genome Illumina paired-end reads (SRA: SRR5123638, SRR2068653, SRR5643140, or ERR701181) to the assembly contigs generated by BWA-MEM (Li 2013). The resulting corrected contigs were reassembled using Flye v2.5 (Kolmogorov *et al.* 2019). For the II×III F1 progeny assemblies, CL13, S27, S21, S26, and D3X1, the assembly contigs were directly subjected to Flye for reassembly without Pilon correction. The final contigs/scaffolds in the *Tg*RH88, *Tg*ME49, and *Tg*CTG assemblies were assigned, ordered, and oriented to chromosomes using ToxoDB-44 genomes as reference (**Table S1**).

Long-read assembly evaluation

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

Whole genome alignment

Whole genome alignments between the long-read assemblies and the reference genomes were performed using MUMmer v4.0.0 (Kurtz *et al.* 2004) and Mauve v2.4.0 (Darling *et al.* 2004). Dotplots were generated using D-Genies (Cabanettes and Klopp 2018). BWA-MEM was used for remapping the corrected reads to the reference genomes, and all SAM files were parsed to sorted BAM files using SAMtools v1.9 (Li *et al.* 2009). Alignments were visualized

using IGV v2.4.15 (Thorvaldsdottir *et al.* 2013).

Structural variant detection

Structural variants between the long-read assemblies and the reference genomes were identified by processing the delta file generated by the MUMmer alignment generator NUCmer with the parameter “show-diff”. In addition, manual curation of structural variants was performed by visual inspection of chromosomal rearrangements based on the whole genome alignments generated using Mauve and using BLASTN to identify and count repetitive loci.

Copy number variant detection

For select the *T. gondii* or *N. caninum* tandemly expanded loci (*ROP5*, *ROP38*, *MIC17*, *MAF1*, *ROP4/7*, and *TSEL8*), all predicted gene sequences were extracted from ToxoDB, and then aligned to the long-read assemblies using BLASTN (Madden *et al.* 1996). Only alignments that showed more than 95% identity and 98% coverage were considered as a match. The type of paralogs was determined by alignment identity, and the number of copies at these loci was estimated by alignment match counts. Only matches that were within a single assembled Nanopore-derived contig were considered for copy number estimates, and the length of the sequence between the upstream of the first match or the downstream of the last match on the genomic coordinate and the edge of the corresponding contig had to be longer than that of the sequence between two adjacent matches.

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

To identify new sequences that filled reference assembly gaps, we aligned the long-read chromosome contigs to the reference assembly chromosomes using NUCmer with the “show-diff -q” parameter. The coordinates of a) sequence expansions and b) unaligned sequences from our *de novo* assembly were determined, and the corresponding sequences were extracted using custom scripts. Repeats in these sequences were detected using Tandem Repeat Finder (TRF) v4.09 (Benson 1999). Simple repeats and low-complexity repeats were predicted using RepeatMasker v4.0.9 (Smit *et al.* 2013-2015) with the RMblast search program against the

apicomplexans database. Gene content of the new sequences in long-read assemblies was assessed using BLASTX against the NCBI non-redundant protein sequences (nr) database and *Toxoplasma gondii* (taxid:5811) was chosen as the organism.

Hi-C Data analysis

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

DATA ACCESS

All raw sequencing data and polished assemblies have been deposited at Genbank under accession numbers XXXX-XXXX (Genbank submission pending).

ACKNOWLEDGEMENTS

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

DISCLOSURE DECLARATION

The authors have no conflicts of interest.

REFERENCES:

Adomako-Ankomah Y, English ED, Danielson JJ, Pernas LF, Parker ML, Boulanger MJ, Dubey JP, Boyle JP. 2016. Host Mitochondrial Association Evolved in the Human Parasite Toxoplasma gondii via Neofunctionalization of a Gene Duplicate. *Genetics* **203**(1): 283-298.

Adomako-Ankomah Y, Wier GM, Borges AL, Wand HE, Boyle JP. 2014. Differential locus expansion distinguishes Toxoplasmatinae species and closely related strains of Toxoplasma gondii. *mBio* **5**(1): e01003-01013.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**(2): 573-580.

Blank ML, Boyle JP. 2018. Effector variation at tandem gene arrays in tissue-dwelling coccidia: who needs antigenic variation anyway? *Current opinion in microbiology* **46**: 86-92.

Brooks CF, Francia ME, Gissot M, Croken MM, Kim K, Striepen B. 2011. Toxoplasma gondii sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proceedings of the National Academy of Sciences of the United States of America* **108**(9): 3767-3772.

Bunnik EM, Venkat A, Shao J, McGovern KE, Batugedara G, Worth D, Prudhomme J, Lapp SA, Andolina C, Ross LS et al. 2019. Comparative 3D genome organization in apicomplexan parasites. *Proceedings of the National Academy of Sciences of the United States of America*.

Burg JL, Grover CM, Pouletty P, Boothroyd JC. 1989. Direct and sensitive detection of a pathogenic protozoan, Toxoplasma gondii, by polymerase chain reaction. *Journal of clinical microbiology* **27**(8): 1787-1792.

Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**: e4958.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved

genomic sequence with rearrangements. *Genome Res* **14**(7): 1394-1403.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics (Oxford, England)* **34**(15): 2666-2669.

Diaz-Viraque F, Pita S, Greif G, Moreira de Souza RdC, Iraola G, Robello C. 2018. Nanopore sequencing significantly improves genome assembly of the eukaryotic protozoan parasite Trypanosoma cruzi. *bioRxiv*: 489534.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333): 92-95.

Edvinsson B, Lappalainen M, Evengard B, Toxoplasmosis ESGf. 2006. Real-time PCR targeting a 529-bp repeat element for diagnosis of toxoplasmosis. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **12**(2): 131-136.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**(3): 186-194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**(3): 175-185.

Fournier T, Gounot JS, Freel K, Cruaud C, Lemainque A, Aury JM, Wincker P, Schacherer J, Friedrich A. 2017. High-Quality de Novo Genome Assembly of the Dekkera bruxellensis Yeast Using Nanopore MinION Sequencing. *G3 (Bethesda, Md)* **7**(10): 3243-3250.

Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ et al. 2008. ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic acids research* **36**(Database issue): D553-556.

Gissot M, Walker R, Delhaye S, Huot L, Hot D, Tomavo S. 2012. Toxoplasma gondii chromodomain protein 1 binds to heterochromatin and colocalises with centromeres and

776 telomeres at the nuclear periphery. *PloS one* **7**(3): e32671.

777 Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny
778 LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome
779 organization. *Nature methods* **9**(10): 999-1003.

780 Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT,
781 Fiddes IT et al. 2018. Nanopore sequencing and assembly of a human genome with
782 ultra-long reads. *Nature biotechnology* **36**(4): 338-345.

783 Jones JL, Holland GN. 2010. Annual burden of ocular toxoplasmosis in the US. *The American*
784 *journal of tropical medicine and hygiene* **82**(3): 464-465.

785 Joynson DH, Wreghitt TG. 2005. *Toxoplasmosis: a comprehensive clinical guide*. Cambridge
786 University Press.

787 Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method
788 employing protein multiple sequence alignments. *Bioinformatics (Oxford, England)* **27**(6):
789 757-763.

790 Khan A, Behnke MS, Dunay IR, White MW, Sibley LD. 2009. Phenotypic and gene expression
791 changes among clonal type I strains of Toxoplasma gondii. *Eukaryotic cell* **8**(12): 1828-
792 1836.

793 Khan A, Shaik JS, Behnke M, Wang Q, Dubey JP, Lorenzi HA, Ajioka JW, Rosenthal BM,
794 Sibley LD. 2014. NextGen sequencing reveals short double crossovers contribute
795 disproportionately to genetic diversity in Toxoplasma gondii. *BMC Genomics* **15**(1):
796 1168.

797 Khan A, Taylor S, Su C, Mackey AJ, Boyle J, Cole R, Glover D, Tang K, Paulsen IT, Berriman
798 M et al. 2005. Composite genome map and recombination parameters derived from
799 three archetypal lineages of Toxoplasma gondii. *Nucleic acids research* **33**(9): 2980-
800 2992.

801 Kim K, Weiss LM. 2004. Toxoplasma gondii: the model apicomplexan. *International journal for*

parasitology **34**(3): 423-432.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**(5): 540-546.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**(5): 722-736.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome biology* **5**(2): R12.

Lapp SA, Geraldo JA, Chien JT, Ay F, Pakala SB, Batugedara G, Humphrey J, Ma Hc, De BJ, Le Roch KG et al. 2018. PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. *Parasitology* **145**(1): 71-84.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:13033997v2 [q-bioGN]*.

-. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**(18): 3094-3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**(16): 2078-2079.

Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, Karamycheva S, Pinney D, Brunk BP, Ajioka JW et al. 2016. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nature Communications* **7**(1): 10147.

Madden TL, Tatusov RL, Zhang J. 1996. Applications of network BLAST server. *Methods in enzymology* **266**: 131-141.

Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P,

828 Aury JM. 2015. Genome assembly using Nanopore-guided long and error-free DNA
829 reads. *BMC Genomics* **16**: 327.
830 Matrajt M, Angel SO, Pszenny V, Guarnera E, Roos DS, Garberi JC. 1999. Arrays of repetitive
831 DNA elements in the largest chromosomes of Toxoplasma gondii. *Genome* **42**(2): 265-
832 269.
833 Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR.
834 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore
835 flow cell. *Nat Commun* **9**(1): 541.
836 Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome
837 assembly evaluation with QUAST-LG. *Bioinformatics (Oxford, England)* **34**(13): i142-
838 i150.
839 Pfefferkorn ER, Pfefferkorn LC. 1976. Toxoplasma gondii: isolation and preliminary
840 characterization of temperature-sensitive mutants. *Experimental parasitology* **39**(3): 365-
841 376.
842 Quick J. 2018. Ultra-long read sequencing protocol for RAD004 V.3.
843 https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-
844 mrxc57n?version_warning=no.
845 Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Könen-Waisman S, Latham SM,
846 Mourier T, Norton R, Quail MA et al. 2012. Comparative Genomics of the Apicomplexan
847 Parasites Toxoplasma gondii and Neospora caninum: Coccidia Differing in Host Range
848 and Transmission Strategy. *PLOS Pathogens* **8**(3): e1002567.
849 Reischl U, Bretagne S, Kruger D, Ernault P, Costa JM. 2003. Comparison of two DNA targets
850 for the diagnosis of Toxoplasmosis by real-time PCR using fluorescence resonance
851 energy transfer hybridization probes. *BMC infectious diseases* **3**: 7.
852 Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors.
853 *Proceedings of the National Academy of Sciences of the United States of America*

74(12): 5463-5467.

Schmidt MH, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Mass J, Pfaff C et al. 2017. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant cell* **29**(10): 2336-2348.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**: 259.

Sibley LD, Ajioka JW. 2008. Population structure of Toxoplasma gondii: clonal expansion driven by infrequent recombination and selective sweeps. *Annual review of microbiology* **62**: 329-351.

Sibley LD, Boothroyd JC. 1992. Construction of a molecular karyotype for Toxoplasma gondii. *Molecular and biochemical parasitology* **51**(2): 291-300.

Smit A, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**(2): 178-192.

Villard O, Candolfi E, Ferguson DJ, Marcellin L, Kien T. 1997. Loss of oral infectivity of tissue cysts of Toxoplasma gondii RH strain to outbred Swiss Webster mice. *International journal for parasitology* **27**(12): 1555-1559.

Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, Wenger AM, Concepcion GT, Kronenberg ZN, Munson KM et al. 2019. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of human genetics*.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS one* **9**(11):

880 e112963.

881 Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV,

882 Zdobnov EM. BUSCO Applications from Quality Assessments to Gene Prediction and

883 Phylogenomics. (1537-1719 (Electronic)).

884