All the analyses described below were run in a Linux environment.

Construction of Contigs

Replicate gzip-compressed sequence reads files (fastq files) were concatenated (using Linux cat command) and submitted in paired-end mode to a locally installed version of Spades (version 3.13.0). The output file (contigs.fasta) was used for downstream analyses.

https://cab.spbu.ru/software/spades/

The command used to run the analysis was as follows:

```
/Shared/Bioinformatics/data/keenhl/spades/SPAdes-3.13.0-
Linux/bin/spades.py \
    -1 /Shared/Bioinformatics/data/keenhl/yao/merged/Yao_R1_001.fastq.gz \
    -2 /Shared/Bioinformatics/data/keenhl/yao/merged/Yao_R2_001.fastq.gz \
    -o output
```

Gene Prediction

The Augustus program was used for gene prediction using the contigs.fasta file previously generated using the Spades program. To run Augustus, a Docker container was created using the Dockerfile instructions on the Augustus GitHub site (below). The Docker container was then converted into a Singularity file using the Singularity build command in order to run on the University of Iowa High Performance Computing cluster.

https://github.com/Gaius-Augustus/Augustus

The command used to run the analysis was as follows:

```
singularity exec ~/privatemodules/augustus/augustus.tar.sif  augustus --
species=toxoplasma contigs.fasta > output.toxoplasma.gff
```

Function Prediction

The InterPro program (version 5.51-85.0) was used for functional prediction using a fasta file of protein sequences extracted from the Augustus output. The protein sequences were extracted from the Augustus output gff file using a custom script written in Perl. To satisfy the dependency of InterPro on Java version 11, a special environment was created with conda and InterPro was ran in this environment. The output from InterPro was processed using code written in R to sort the different analyses and create an integrated Excel file.

https://interproscan-docs.readthedocs.io/en/latest/

The commands used to run the analysis were as follows where openjdk_11 is the name of the conda environment and chao.fasta is a fasta file containing predicted protein sequence from Augustus:

```
conda activate openjdk_11

./interproscan.sh -i chao.fasta

conda deactivate
```