Long-Read Genome Assembly and Gene Model Annotations for the Rodent Malaria Parasite *Plasmodium yoelii* 17XNL

Mitchell J. Godin[1]*, Aswathy Sebastian[2]*, Istvan Albert[1,2#], Scott E. Lindner[1#]

* Equal Contributions, # Co-Corresponding Authors

1. Department of Biochemistry and Molecular Biology, The Huck Center for Malaria Research, The Center for Eukaryotic Gene Regulation, Pennsylvania State University, University Park, PA, 16802

2. Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, 16802

Corresponding Authors:

Istvan Albert, Ph.D.
206C Life Sciences Building
University Park, PA 16802
P: +1.814.865.2281
E: iua1@psu.edu
ORCID: 0000-0001-8366-984X

Scott E. Lindner, Ph.D.
W230B Millennium Science Complex
University Park, PA 16802
P: +1.814.867.4062
E: Scott.Lindner@psu.edu
ORCID: 0000-0003-1799-3726

Keywords: *Plasmodium yoelii* 17XNL, malaria, genome assembly, gene model

Abstract

Malaria causes over 200 million infections and over 600 thousand fatalities each year, with most cases attributed to a human-infectious *Plasmodium* species, *Plasmodium falciparum*. Many rodent-infectious *Plasmodium* species, like *Plasmodium berghei*, *Plasmodium chabaudi*, and *Plasmodium yoelii,* have been used as genetically tractable model species that can expedite studies of this pathogen. In particular, *P. yoelii* is an especially good model for investigating the mosquito and liver stages of parasite development because key attributes closely resemble those of *P. falciparum*. Because of its importance to malaria research, in 2002 the 17XNL strain of *P. yoelii* was the first rodent malaria parasite to be sequenced. While sequencing and assembling this genome was a breakthrough effort, the final assembly consisted of >5000 contiguous sequences that impacted the creation of annotated gene models. While other important rodent malaria parasite genomes have been sequenced and annotated since then, including the related *P. yoelii* 17X strain, the 17XNL strain has not. As a result, genomic data for 17X has become the *de facto* reference genome for the 17XNL strain while leaving open questions surrounding possible differences between the 17XNL and 17X genomes. In this work, we present a high-quality genome assembly for *P. yoelii* 17XNL using HiFi PacBio long-read DNA sequencing. In addition, we use Nanopore long-read direct RNA-seq and Illumina short-read sequencing of mixed blood stages to create complete gene models that include not only coding sequences but also alternate transcript isoforms, and 5’ and 3’ UTR designations. A comparison of the 17X and this new 17XNL assembly revealed biologically meaningful differences between the strains due to the presence of coding sequence variants. Taken together, our work provides a new genomic and gene expression framework for studies with this commonly used rodent malaria model species.

Introduction

Malaria remains a major global health burden (WHO Malaria Report 2022, (1)), with most of the 600,000 fatalities resulting from infection by human-infectious *Plasmodium falciparum*. The use of rodent-infectious model species has been instrumental to better understand those species that cause human disease due to high levels of genetic and physiological conservation across species (2). Researchers have routinely used these rodent model species, such as *P. yoelii*, *P. berghei*, and *P. chabaudi*, to investigate the entire *Plasmodium* life cycle, as genetic manipulations have long been rapid and rigorous in these species (2). We and others study *P. yoelii*, which is an especially good model for the mosquito and liver stages of *P. falciparum* parasite development (2). This is partly because *P. yoelii* mosquito stage parasites develop at a similar pace as do those of *P. falciparum*, and their sporozoites are less promiscuous than *P. berghei* sporozoites (2). Because of this, many studies of genetically attenuated parasite (GAP) vaccine candidates based upon sporozoites have recently included the use of *P. yoelii* as a pre-clinical model system (3). In support of this, large-scale analyses of gene expression of *P. yoelii* now match those available for *P. berghei* in many ways (4-9). For these reasons, *P. yoelii* has been an important malaria parasite used as a proxy for *P. falciparum* in pre-clinical and discovery phase studies.

Intuitively, genetic studies of any species are best conducted with accurate genome assemblies and gene models. Therefore, several species of *Plasmodium* parasites were the subject of early whole-genome sequencing efforts in the late 1990s and early 2000s (10, 11). This work provided a genome assembly of the human-infectious *P. falciparum* parasite with 14 nuclear chromosomes and the two organellar genomes of its mitochondrion and apicoplast (11). In addition, gene models for *P. falciparum* were annotated with introns/exons, with further improvements establishing 5’/3’ untranslated regions (UTRs) and transcript isoforms (12, 13). Similarly, the rodent-infectious *Plasmodium berghei* ANKA parasite was originally sequenced in 2005, resulting in a genome assembly with 7,497 contiguous sequences (contigs) that were later reduced to 16 contigs with a hybrid Illumina and 454 sequencing approach in 2014, and then further refined using PacBio sequencing in 2016 (14-16). Prior to this, the non-lethal *P. yoelii* 17XNL strain was the first rodent malaria parasite sequenced in 2002, which used ABI3700 sequencers and yielded a genome assembly of over 5,000 contigs (10). The *P. yoelii* 17X strain, from which 17XNL was derived, was sequenced in 2014 alongside PbANKA using the same Illumina and 454 sequencing approach to similarly establish a 16 contig genome assembly (15). With the advent of more accurate long-read sequencing technologies, there has been a renewed interest in sequencing the *Plasmodium* genomes and transcriptomes, including those of another *P. yoelii* strain, PyN67, which has been used to study genetic polymorphisms and drug responses (17). In addition, the genomes of other apicomplexan parasites, such as *Cryptosporidium* and *Babesia* species, have now been established using a combination of long-read Nanopore sequencing and short-read Illumina sequencing (18-20).

Although their genomes have been updated and are conveniently provided on PlasmoDB.org, Py17X and PbANKA have gene models that largely reflect the coding sequences, but not their UTRs despite the availability of RNA-seq data that could be used to approximate them (4-9, 15, 21-25). Finally, while the 17XNL strain of *P. yoelii* remains a highly used laboratory strain worldwide, its reference genome and gene models have not been revisited since 2002, and thus its genome assembly and gene models remain highly fragmented and incomplete. As a result, most researchers use the genome assembly and gene models of the related *P. yoelii* 17X strain as a proxy when working with the 17XNL strain and must operate under the assumption that the genomes of the two strains are effectively the same. However, this prompts a few important questions. How similar are the 17X and 17XNL strains? In what ways are they truly suitable proxies for one another? Given the state of the 17XNL genome assembly and the limited gene models available for both strains, these questions could not be accurately addressed. However, these kinds of questions can now be more rigorously addressed with the inclusion of long-read DNA sequencing. The long sequence reads produced by PacBio and Nanopore approaches better facilitate the scaffolding of long, contiguous sequences in a *de novo* assembly, even for complex genomes that have extreme AT-content and/or high degrees of repetitiveness, such as found with *Plasmodium* species (26, 27). Additionally, by combining long-read and short-read sequencing, the strengths of each can be used to polish the assembly to reduce systematic errors introduced by each of the different methodologies.

Therefore, here we have created a high-quality reference genome and gene model annotation for the *P. yoelii* 17XNL strain that we have used to address these outstanding questions. We utilized HiFi PacBio DNA-seq to create a Py17XNL reference genome with 16 high confidence/high accuracy contigs. Even without any polishing efforts, this approach outperformed a parallel effort using a hybrid Nanopore long-read DNA-seq/Illumina short-read DNA-seq method by several key metrics, including its assembly quality and the reduction of gaps. Furthermore, we created gene annotations for genes transcribed in asexual and sexual blood stages using a combination of Nanopore direct RNA-seq and our pre-existing Illumina RNA-seq datasets. These annotations include definitions of introns, exons, 5', and 3' UTRs, and transcript isoforms expressed in asexual and sexual blood stages. Using these data, we compared the genomic variance between the Py17XNL and Py17X strains to gain insight into the differences between the two strains and identified that most sequence variants reside in intergenic regions, whilst variation in the coding sequence of a select few genes could result in meaningful changes in Py17XNL parasite biology.

Results

A Comparison of Genome Assembly Approaches: PacBio HiFi vs. Nanopore/Illumina Hybrid Sequencing

P. yoelii 17XNL remains a commonly studied rodent malaria strain. Yet, its genome assembly remains highly fragmented and consists of over 5000 contigs as generated in 2002 (10). Consequently, most researchers use the reference genome of the related Py17X strain as a substitute for Py17XNL without knowing how appropriate it is to use it as a genomic proxy (15). To resolve these questions, we created a high-quality genome assembly of *P. yoelii* 17XNL Clone 1.1 obtained from BEI Resources, which is the common origin of this strain of parasites for many laboratories. Because several sequencing methodologies are now commonly used to assemble whole genomes, we used Nanopore, PacBio, and Illumina sequencing with DNA- or RNA-based libraries to determine an optimal approach to create a genome assembly with associated gene models for *P. yoelii* 17XNL. Nanopore ligation-based long-read DNA sequencing is currently favored by many researchers as it can provide extremely long sequence reads, resolve long stretches of repetitive regions, and assemble long structural variants in the genome (26, 27). HiFi PacBio DNA sequencing provides very high accuracy due to the sequencing of circular consensus sequences (ccs) of ~10kb DNA fragments, providing a middle ground between the sequencing sizes provided by Illumina and Nanopore sequencing (27). We explored several data analysis protocols for combining data from different platforms to optimize this genome assembly. As widely available Nanopore sequencing chemistries (Q10) yield a systematic error, Nanopore data are often paired with Illumina data to provide error correction. The sequencing error rates from Illumina are typically above 99.9% and can be used with polishing algorithms to identify errors in assemblies that were produced with long, noisy reads (18, 19). A detailed outline of our experimental methods is included in Supplemental Figure 1. Briefly, swiss webster outbred mice were infected with Py17XNL Clone 1.1 parasites that had been passaged only once following receipt from BEI Resources to create a genome assembly reflective of the current stocks available in the depository. Upon reaching 1-3% parasitemia, mice were euthanized, white cells were depleted by cellulose, and red blood cells (RBCs) were lysed by saponin. The parasite pellets were used to produce high molecular weight genomic DNA using the NEB Monarch HMW DNA Extraction Kit for Cells and Blood as previously described (28). DNA purity, quantity, and fragment lengths were determined to all be high quality by NanoDrop, Qubit, and TapeStation measurements, respectively (Supplementary Table 1, Supplementary Figure 2)). This approach yielded DNA fragments of higher quality and higher molecular weight than the Qiagen QIAamp DNA Blood Mini Kit that is routinely used in our laboratory and in others. Matched gDNA samples were sequenced on a Nanopore MinION R9.4.1 flow cell using the ligation sequencing kit, as well as on an Illumina NextSeq 550 using the Illumina DNA PCR-Free kit. In parallel, Py17XNL HMW gDNA was sequenced on a PacBio Sequel using a PacBio SMRT cell. The raw reads from the PacBio sequencing run were converted into circular consensus sequences using the CCS algorithm.

To assess the quality of both Nanopore sequencing runs, we utilized Nanoplot, a quality control plotting suite specifically for long-read sequencing data (Supplemental Figure 3, Supplemental Table 2)) (29). The Nanopore sequencing runs for both replicate one and two resulted in an overall average read length of 16,706 bases, with an average Qscore of 11.3 (Supplemental Table 2). During this study, an improved high-accuracy base calling algorithm from Nanopore was released, which we also tested to see if it could improve our read quality. Upon re-basecalling the fast5 files, we saw a considerable increase in Qscore, from 11.3 to 14.1, even without the quality score filter that is imposed with the default, fast basecalling algorithm (Supplemental Figure 3, Supplemental Table 2). Despite this improvement in quality, there were no significant differences in the mean read length or the throughput (Supplemental Table 2). Using PacBio ccs reads, we saw an improvement in accuracy to an average Qscore of 36.3 (Figure 1A). The biggest difference came with throughput, which increased to 1,660,222,360 bases from 707,945,539 bases whilst still maintaining an average read length of 5,712.7 bases (Figure 1B, Supplemental Table 2).

We generated genome assemblies from both long-read datasets with the bioinformatic workflows described in Figure 2. To create the Nanopore/Illumina hybrid genome assembly, we assembled the Py17XNL Nanopore data using Flye (30) and scaffolded the contigs using the Py17X genome as a guide in conjunction with the RagTag scaffolding program (30, 31). Finally, we layered error correction onto it in a multi-step approach, first using nextpolish, followed by multiple rounds of consensus generation based on Illumina data alignment and variant calling (Figure 2) (32). Through this process, we were able to reduce the number of contigs down to 16, but at the cost of covering less of the genome (95%) and introducing 34 misassembles (Figure 1C) as defined by the assembly evaluator program Quast (33). The PacBio-based genome assembly was generated with the HiCanu program (34), which produced a *de novo* genome assembly with 132 contigs (Figure 2). The resulting contigs were filtered to contain only the target species by aligning them against the Py17X genome using minimap2 (35). Contigs that had a primary alignment length of >2% of the 17X reference chromosome were assigned the matching chromosome names. A consensus genome was then created by aligning these contigs with the 17X reference genome and filling in the missing genomic regions, mainly chromosomal ends. This resulted in a final assembly of 23.08 Mb with 16 contigs (Figure 1C, Table 1). We have adopted the higher quality PacBio-based genome assembly for the *P. yoelii* 17XNL strain for the rest of our analyses and for provision to the community on PlasmoDB.org, which we term Py17XNL_2 to distinguish it from the original genome assembly (Py17XNL_1) (10, 36). However, as both the hybrid Nanopore/Illumina and PacBio assemblies are potentially valuable to our research community, both assemblies have been publicly deposited in NCBI.

Nanopore Direct RNA sequencing provides new information to pre-existing gene models

We also set out to create more comprehensive gene models to increase the utility of the new Py17XNL_2 genome assembly for the P. yoelii 17XNL strain. In the currently available gene models for *P.*

berghei (ANKA) and *P. yoelii* (17X, 17XNL) on PlasmoDB, only the coding sequences of genes are provided with no designation of untranslated regions (UTRs), and little information is provided about alternatively spliced transcripts. We generated gene models that provide complete transcript information, including start and stop codons, transcription start and stop sites, and UTRs. Experimentally, we performed Nanopore direct RNA-seq in biological duplicate to generate long sequence reads of asexual and sexual blood stage transcripts. Briefly, total RNA was extracted from parasitized mouse blood to create an RNA-seq library that was sequenced with the Nanopore Direct RNA-Sequencing Kit (Supplemental Figure 1, Supplemental Figure 4, Supplemental Table 1). These direct RNA-sequencing reads were quality controlled using Nanoplot with the same parameters described for Nanopore ligation DNA-sequencing for both “Fast” and “High Accuracy” basecalling approaches (Supplemental Figure 5) (29). In total, 429,888,068 bases were sequenced after combining the replicates, with an average Qscore of 12 after high accuracy basecalling, which again outperformed the fast basecalling approach (Supplemental Figure 5, Supplemental Table 2)). The mean read length across replicates was 858 bases, with the longest read being 8,789 bases (Supplemental Figure 5, Supplemental Table 2).

Gene models were created with two alternative methodologies using Nanopore direct RNA-seq long reads alone or in combination with our previously published Illumina short-read RNA-seq of mixed asexual and sexual blood stage parasites (Figure 2) (37). These parallel approaches are both informative, given the strengths and limitations of both sequencing techniques. Nanopore direct RNA sequencing provides information that allows us to identify long/full-length sequencing reads that initiate at the 3’ end of mRNAs (38). However, when the full-length mRNA is not sequenced, less information is provided for the 5’ end (38). This limitation is remedied by the strong depth and breadth of sequencing coverage provided via Illumina sequencing. For both approaches, Nanopore RNA-seq reads were aligned to our Py17XNL_2 genome using minimap2 (35). For gene models created with both Nanopore and Illumina

RNA-seq data, in parallel, the Illumina short reads were aligned to the Py17XNL_2 genome using Hisat2 (31375807). To create the gene models and assign gene names/descriptions, we used Braker2, as well as reciprocal blast searches using the blastp program of the BLAST suite (39, 40). The Nanopore-only approach helped us identify 5,683 genes, 5,828 mRNAs, 66 tRNAs, and 40 rRNAs. Using the Nanopore/Illumina hybrid approach, we found 6,077 genes, 7,047 mRNAs, 66 tRNAs, and 40 rRNAs (Table 1). Gene models that were generated using both Nanopore and Illumina reads more closely matched the anticipated UTR length that was defined in recent *Plasmodium falciparum* transcriptomics data (13). A representative example of this more comprehensive gene model is illustrated in Figure 3A.

Nanopore direct RNA-seq initiates at the 3' end due to the use of poly(dT) sequencing primers. As a result, significantly higher coverage was obtained for the 3' UTRs than for 5' UTRs (Supplemental Table 3). The higher coverage allows us to further analyze the 3' UTR length distribution for *P. yoelii* 17XNL, which is of interest as *cis*-regulatory elements are often found in this portion of eukaryotic mRNAs (41). The majority of reads have a UTR length between 100 and 200 bp, with a mean length of 364 bp. The largest UTR reported for the H2B.Z histone variant mRNA, with 1994 nt (Figure 3B, Supplemental Table 3). Compared to the most up-to-date *P. falciparum* transcriptome, which used DAFT-Seq to resolve UTRs, P. yoelii 17XNL's 3' UTRs appear slightly shorter on average (Figure 3C) (13).

Comparison between Py17XNL_2 and reference genomes demonstrates the completeness of the assembly

Using the Py17XNL_2 genome assembly and associated gene models, we compared our results to the original Py17XNL genome (Py17XNL_1) and the Py17X reference genome. As anticipated, there was a substantial reduction in the number of gaps/misassembles and greater genome coverage when comparing Py17XNL_2 vs. Py17XNL_1 (Table 1). Although our Nanopore/Illumina-based genome assembly (Py17XNL Nanopore) contained the same number of contigs as the PacBio-based Py17XNL_2 assembly, the significantly fewer misassemblies generated in the PacBio-based assembly provided a more accurate reference genome for future research uses. Additionally, our final PacBio-based assembly (Py17XNL_2) closely resembles that of Py17X, which was also created using recently developed sequencing technologies (15). However, when compared to Py17X, the Py17XNL_2 assembly has lower coverage of repetitive sequences at the sub-telomeric regions, which precluded us from robustly assembling these regions for some chromosomes, requiring consensus generation based on alignments to the Py17X reference genome. Similarly, the 6,086 new gene annotations more accurately represent the anticipated number of genes for Py17XNL and more closely match those annotated in other *Plasmodium* species (Table 1). Moreover, these new gene models include both coding sequences, UTRs, and transcript isoforms, which are lacking in the provided gene models currently available on PlasmoDB for this specific species. In addition to these assessment metrics, we determined the completeness of the reference genome based on marker genes. To quantify this, we used a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis that detects whether a predefined set of single-copy marker genes in the *Plasmodium* lineage are present in these data (Figure 4) (42). This BUSCO dataset contains 3642 BUSCO groups from 23 different species, including *P. falciparum* 3D7, *P. yoelii* 17XNL, *P. vivax, P. berghei* ANKA, *P. chabaudi,* and others. From this search, 3556/3642 (97.6%) of complete and single copy BUSCOs were found to be present, indicating that this genome assembly and gene annotation has a high level of completeness (Figure 4).

<u>Variation between the Py17XNL and Py17X reference genomes primarily resides in the intergenic regions and the ends of chromosomes</u>

As Py17X is commonly used as an interchangeable proxy genome for Py17XNL, we sought to determine what similarities and differences exist between the strains and how the differences may impact genetic studies. We performed chromosome-wide alignments between the Py17X and

Py17XNL_2 genomic builds using the minimap2 (35) program and assessed the genome-wide variants with paftools. We observed extensive linear agreement between the two strains, with 99.9% of the Py17X genome matching with the Py17XNL_2 genomic build (Figure 5A). We did not detect any large structural variation between the strains, a finding also supported through our alternative Py17XNL_2 Nanopore/Illumina genome build. At the same time, we also identified a total of 1,955 potential single nucleotide/short variants across the two strains, the majority (62%) of which were found in intergenic regions (Figure 5B). We found that the apicoplast genome was identical between strains, whereas the Py17XNL_2 mitochondrial genome has a 127 bp deletion in the middle of its sequence. Compared to Py17X, the deletion is located in the intergenic region between *cox1* (PY17X_MIT00800) and a ribosomal RNA fragment annotated as PY17X_MIT00700. Together, we conclude that while these two strains are highly similar, there are sequence differences that may be functionally relevant.

To further determine the potential impacts of these genome variants, we characterized the position of variants with respect to nearby genes and, when applicable, determined the specific DNA and amino acid changes that would result from the change. Most variants were found to be located in intergenic regions and were characterized as single base pair indels (Figure 6, Figure 7A,B). Of the 334 variants that fell within coding regions, we characterized the changes in nucleotide and amino acid protein composition of the encoded proteins (representative examples are provided in Figure 6, nucleotide and amino acid level changes are provided in Supplemental Table 4). Due to the over representation of single bp indels, the majority of amino acid changes lead to frameshifts (Figure 7C). Most of these frameshifts took place in genes that were unnamed with an unknown function, requiring further investigation to determine the biological impacts of these differences.

To further interrogate those variants that occurred in well-characterized genes, we manually curated the results and verified the variant calls via various quality measures. We checked if the variant had sufficient PacBio ccs read support (80% of reads support the variant with a minimum of 5x coverage at the region), and when possible, also determined if additional Nanopore/Illumina DNA and RNA sequencing reads supported the variant (80% of reads support the variant with a minimum of 5x coverage for Illumina sequencing and 2x coverage for Nanopore sequencing). Through manual curation, a substantial number of variants had support from at least three sequencing methodologies. Due to the strict thresholds of this variant calling process, some sequencing methods did not capture the variant sufficiently enough to provide support, typically due to a lack of coverage at the position of the variant. An example of this occurring is with CSP, which had a large deletion that was adequately supported by PacBio and Nanopore DNA-seq data (Figure 6). Illumina DNA-seq reads, which should capture this variant due to their high accuracy, instead have a complete loss of coverage, with only one read correctly mapping to this repetitive region (Figure 6, Bottom Panel). As a result, we encourage the use of long-read sequencing platforms to identify variants that may be missed when using Illumina sequencing.

Based upon these criteria, we identified if these genes were expressed in asexual/sexual blood stages due to sequencing support from either Illumina or Nanopore direct RNA-seq and created separate variant lists accordingly (Supplemental Table 4). Finally, we filtered out genes with no annotated gene name and those that belong to a variable gene family (fam/pir gene families) (Figure 6). After this filtering, we focused our analyses on the remaining 13 blood stage-expressed genes (Table 2). Although the biological implications of the differences between Py17X and Py17XNL will need further experimental validation, many variants could have interesting impacts. One such example is *ap2-sp*, which has both synonymous and missense variants between the AT-hook and AP2 domain (43, 44). AP2-SP is an ApiAP2 transcription factor with many target genes that are expressed specifically in the sporozoite stage of the *Plasmodium* life cycle (43, 45, 46). It has also been shown that disruption of this gene results in the loss of sporozoite formation entirely in the related *P. berghei* parasite and has important activities in blood stages in *P. falciparum* (43, 45-47). Another affected gene is *pk4*, which encodes an essential eIF2α kinase related enzyme and contains changes in its non-cytoplasmic domain as determined by InterPro domain predictions across 17X and 17XNL strains (48-51). Further study of these genes and several other candidates is warranted to understand the biological role these variants may play across strains.

Among the non-blood stage expressed genes, *trap, lisp2,* and *csp* all had variants in their coding sequences when comparing 17XNL to 17X. Of these, the most notable one is a large in-frame deletion within the repeat region of CSP, leading to the loss of six of the repeating units of D/PQGPGA in Py17XNL (Supplemental Figure 6). Similarly, the YM strain of *P. yoelii* is even shorter and lacks an additional repeating unit compared to 17XNL (Supplemental Figure 6) (15). In *P. berghei*, it was found that 25% of this repeat region could be eliminated before impacting parasite development, which is approximately the length reduction observed in 17XNL and YM as compared to 17X (52). Therefore, this may reflect a minimum repeat length for CSP functions.

Discussion

Here we have created a high-quality genome assembly with experimentally validated gene models for the commonly used 17XNL strain of the *P. yoelii* malaria parasite species. We envision this will be an important resource to the malaria research community, as it provides a much-needed update to the Py17XNL_1 reference genome, which was among the first to be completed in the early days of the genomics era 20 years ago (10). By directly comparing the strengths and genome assemblies created from either PacBio HiFi sequencing or a combination of Nanopore DNA-seq and Illumina DNA-seq, we identified that while the hybrid Nanopore/Illumina approach yielded a robust genome assembly, the PacBio HiFi-based assembly consisted of fewer misassembles and covered a greater fraction of the genome. Therefore, we have chosen the PacBio-based genome assembly as our new working reference genome for *P. yoelii* 17XNL strain, which we have designated as Py17XNL_2 within this study. Our findings align with many recent studies conducted to improve the reference information on *Plasmodium* species, which also utilized an exclusively PacBio-based approach (12, 14, 17, 53). We also deemed it important to conduct both approaches to leverage the strengths of Nanopore sequencing, which permit greater detection of large-scale structural variants in the genome as compared to approaches with shorter read lengths (26). The strengths of Nanopore sequencing have also been leveraged by other sequencing efforts, most notably the recent telomere-to-telomere sequencing effort of the human genome that used ultra-long read approaches (54). During this study, advances in Nanopore basecalling software were made that enabled more accurate sequencing without the need for re-sequencing or new hardware. We therefore directly compared the previous “fast” vs. new “high accuracy” basecalling algorithms and observed a substantial increase in Qscores associated with the same DNA and RNA sequencing data (Supplemental Table 2). However, even with the use of high accuracy basecalling, PacBio data still enabled the most accurate Py17XNL genome assembly and covered the greatest fraction of the genome.

To provide an even more useful genome reference, here we also provide new gene annotations for the 17XNL strain of *P. yoelii* to facilitate more reliable forward and reverse genetic studies of this key model malaria species. Regardless of the rodent malaria RNA-seq studies that have been performed, gene models available on PlasmoDB for *P. berghei* and *P. yoelii* only consist of their putative coding sequences. Here we have now added experimentally validated information on alternatively spliced transcripts and untranslated regions (UTRs) of Py17XNL blood stage-expressed genes. To date, the only other comparable efforts in our field have been applied to *P. falciparum* with a focus on either identifying alternatively spliced transcripts or experimentally defining and annotating long noncoding RNAs (lncRNAs) (55-57). Additionally, because Nanopore direct RNA-seq reads initiate at the 3' end of mRNAs and progress toward the 5' end, it is also strong-suited in providing information about the 3' UTRs of a population of mRNAs. From this, we created both a Nanopore-only and a Nanopore/Illumina hybrid gene model annotation that can both be useful to researchers depending on the questions they are pursuing. We are therefore providing both gene model files as resources to our community. These gene models include well-defined 3'UTRs for Py17XNL that are in agreement with the length distribution of those described for *P. falciparum* (13). Due to the strengths of this approach, we anticipate that Nanopore direct RNA sequencing will become a useful tool for future work on *Plasmodium* parasites, especially as sequencing chemistry and basecalling algorithms improve.

With this greatly improved Py17XNL reference genome, we also were able to critically analyze genomic variation across the 17X and 17XNL strains of *P. yoelii*. As it is currently common practice to use Py17X as a proxy genome for Py17XNL for genomic studies, we thought it was important to begin addressing whether biologically relevant differences were present that would impact such efforts. By aligning the two genomes, we saw that there was an excellent linear agreement between them, with most variation taking place in intergenic regions. In total, there were 1,955 variants across the entire sequence, with 334 of those being in the coding sequence of genes. Most of these variants were single bp indels that most likely accounted for the overrepresentation of frameshift variants in the respective amino acid sequence. Upon further analysis of these variants, some interesting questions arose regarding the biological implications that these changes could have. Specific examples of genes with impactful variants include the ApiAP2 transcription factor AP2-SP and PK4, which are essential for *Plasmodium* development and warrant follow-up studies (43, 45, 46, 48, 50, 51). Aside from these blood stage-expressed genes, it is also important to note the large-scale differences between the 17X, 17XNL, and YM strains of *P. yoelii* in the central repeat of CSP, which are 150, 114, and 108 amino acids long, respectively (Supplemental Figure 6). The in-frame deletions result in truncations of entire six amino acid repeating units of D/PQGPGA, with 17XNL having six fewer units and YM having seven fewer than 17X. In *P. berghei*, it was found that 25% of this repeat region could be eliminated before impacting parasite development, which reflects the approximate reduction in repeat length in 17XNL and YM strains as compared to 17X (52). We anticipate that this may reflect a minimum repeat length that is applicable to both highly related species. This is indirectly corroborated by the absence of any reports that have documented significant differences in sporozoite development, functions, or transmissibility between the 17X and 17XNL strains. Additionally, this particular variant was identified in both Nanopore and PacBio long-read DNA-sequencing datasets, with Illumina short-read sequencing lacking coverage at this site to accurately identify this deletion (Figure 6). This highlights the utility of long-read sequencing technologies to resolve highly repetitive genomes.

This *P. yoelii* 17XNL_2 reference genome and its more comprehensive gene annotations provide a resource that we believe will be helpful to the rodent malaria research community. We stress that while most genes are identical between the 17X and 17XNL strains, there is appreciable genomic variance in some important genes that should be considered when conducting genomic studies. Therefore, we conclude that for many efforts, 17X is a suitable genomic proxy for 17XNL, but caution against it for genes where variance exists, such as *csp, trap, lisp2, ap2-sp, pk4*, and others. Given the improvements for both the Py17XNL_2 genome assembly and gene models presented here, we would instead encourage their adoption as the working reference genome and gene annotation source for studies of *P. yoelii* 17XNL.

Materials and Methods

Animal Experiments Statement

All animal care strictly followed the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) guidelines and was approved by the Pennsylvania State University Institutional Animal Care and Use Committee (IACUC# PRAMS201342678). All procedures involving vertebrate animals were conducted in strict accordance with the recommendations in the Guide for Care and Use of Laboratory Animals of the National Institutes of Health with approved Office for Laboratory Animal Welfare (OLAW) assurance.

Experimental Animals

Six-to-eight-week-old female swiss webster mice from Envigo were used for all experiments in this work.

Parasite Preparation and Isolation

Mice infected with wild-type Py17XNL Clone 1.1 parasites from BEI Resources until a parasitemia between 1-3% was reached. Approximately 1 mL of blood was collected from each euthanized mouse, which was then added to 5 mL of heparinized (200 U) 1X PBS to prevent coagulation. The infected blood was spun and the serum was aspirated to isolate the red blood cells (RBCs). Cells were resuspended in 10 mL 1X PBS and then passed through a cellulose column (Sigma #C6288) to remove mouse leukocytes. The RBCs were then lysed in 0.1% w/v saponin in 1X PBS for 5 minutes at room temperature, and parasite pellets were subsequently washed in 10 mL 1X PBS.

gDNA Preparation

All gDNA samples used for Nanopore, Illumina, and PacBio sequencing were prepared using the NEB Monarch HMW DNA Extraction Kit for Cells and Blood (NEB #T3050) using the manufacturer's protocol for fresh blood with slight modifications as we have previously described (28). Briefly, the saponin lysed parasite pellet was resuspended in 150 µL of Nuclei Prep Buffer containing RNase A. After resuspension of the pellet, 150 µL of Nuclei Lysis Buffer containing Proteinase K was added and mixed by inversion. The sample was then placed in a thermal mixer at 56°C with an agitation speed of 1500 rpm for 10 minutes. Next, 75 µL of precipitation enhancer was added and mixed by inversion. Two DNA capture beads were added to the tube, along with 275 µL of isopropanol. The sample was then mixed 30 times with manual, slow, end-over-end inversions to ensure the gDNA stuck to the capture beads. The supernatant was removed, and the beads were washed twice with 500 µL of gDNA wash buffer. Subsequently, 100 µL of elution buffer II was added and the sample was incubated for five minutes at 56°C in a thermal mixer with agitation at 300 rpm. The beads were added to a bead retainer in an Eppendorf tube, and the sample was spun down for 30 seconds at 12,000 *xg*. All samples were stored at 4°C to minimize shearing from freeze-thaw cycles. Fresh gDNA samples were made for replicate 1 and replicate 2 for Nanopore sequencing and the sole sample for PacBio. The same gDNA samples used for Nanopore replicates 1 and 2 were used for Illumina DNA sequencing replicates 1 and 2. Sample concentration and purity were assessed via Qubit and Nanodrop, respectively (Thermo Fisher Scientific® Nanodrop® 2000 and Qubit® instruments with the Qubit dsDNA BR Assay Kit (Cat #Q32853)). Fragment length was assessed using an Agilent Technologies® TapeStation® 4200 system with Genomic DNA ScreenTapes (Cat #5067-5366 and 5067-5365).

RNA Preparation

RNA samples were prepared from two biological replicates for Nanopore direct RNA sequencing. RNA samples were produced using the Qiagen RNeasy kit using the manufacturer's protocol with slight modifications to improve yield (Cat # 74104). Briefly, 350 µL of Buffer RLT was added to resuspend the parasite pellet. The sample was passed through a 20-gauge needle five times and put back into the same microfuge tube. Next, 350 µL of 70% ethanol was then added and mixed by pipetting using wide-bore pipettes. The sample was then added to the spin column and was centrifuged for 15 seconds at 8,000 *xg*. The column was washed twice with 500 µL RPE buffer and was again centrifuged for 15-60 seconds at 8,000 *xg*. Residual ethanol was removed by a final spin at these parameters. RNA was eluted from the column into a fresh microfuge tube with 30 µL of DEPC-treated water. The sample was incubated for 15 minutes at room temperature to improve recovery yield. The sample was then collected by centrifugation for 1 minute at 8,000 *xg*. A second elution with 30 µL DEPC-treated water was performed as above to improve yield. To eliminate contaminating DNA, a Dnase I digestion was performed with slight modifications to the Sigma #AMPD1 technical bulletin. Briefly, 8 µL of the prepared RNA was mixed with 1 µL 10X Reaction buffer and 1 µL Dnase I, Amplification Grade, 1 unit/µL (Cat # AMPD1-KT). The sample was gently mixed and incubated at room temperature for 30 minutes. The digestion was terminated by the addition of 1 µL stop solution, followed by heat inactivation of the Dnase I. RNA was precipitated with ethanol by adding 0.1 volume of 3M sodium acetate pH5.5@RT, four volumes of reagent grade 200 proof ethanol, and 0.5 µL 20mg/ml glycogen. The solution was allowed to precipitate overnight at -80°C. The solution was then spun down at 4°C at 12,000 *xg* for 10 minutes. The supernatant was aspirated and 1 mL of 70% ethanol was added to wash the pellet. The pellet was spun down as above and the supernatant was aspirated. The pellet was then allowed to air dry with the tube inverted on a Kimwipe for 10 minutes. Sample concentration and purity were assessed via Qubit and Nanodrop, respectively (Thermo Fisher Scientific® Nanodrop® 2000 and Qubit® instruments with the Qubit dsDNA BR Assay Kit (Cat #Q32853)). RNA integrity was tested using an Agilent 2100 Bioanalyzer.

Nanopore Ligation-based DNA Sequencing

DNA sequencing of Nanopore replicates 1 and 2 was performed using the SQK-LSK110 Ligation sequencing kit using the manufacturer's protocol. Genomic DNA (~1 µg as measured by Qubit) was sequenced on an R9.4.1 (Cat # FLO-MIN106D) flow cell for 24 hours, washing between samples as per manufacturer's recommendations (EXP-WSH003).

Nanopore Direct RNA Sequencing

RNA sequencing of Nanopore replicates 1 and 2 was performed using the SQK-RNA0002 Direct RNA-sequencing kit from Oxford Nanopore Technologies using 500ng RNA. All sequencing was performed on an R9.4.1 flow cell for 24 hours, washing between samples as per manufacturer's recommendations (EXP-WSH004).

Illumina DNA Sequencing

Illumina DNA sequencing libraries were created using the Illumina DNA PCR-Free Kit with 100 ng of total input (Cat # 20041794). Illumina libraries were sequenced on a NextSeq 550 Mid Output 150x150 paired-end sequencing run.

PacBio Sequencing

PacBio libraries were created using the PacBio SMRTbell Express Template Prep kit 2.0 (Cat # TPK 2.0) using an input of 2 µg gDNA that was sheared with the Covaris g-TUBE to an average fragment length of 10kb (Cat # 520079). The library was sequenced on a PacBio Sequel using a SMRT Cell 1M v3 LR at a 10 pM library loading concentration with a 2-hour pre-extension time and a 20-hour movie time (Cat # 101-531-000).

Data Analysis

High-quality HiFi reads were extracted from PacBio sequencing data requiring a minimum of 3 full passes in CCS command (v6.0.0) (31562484). The HiFi reads were *de novo* assembled using HiCanu, specifying a genome size of 23 Mb. The resulting 132 contigs were aligned to the Py17X reference genome using minimap2 (35), and all small contigs that had <2% alignment with a 17X chromosome were filtered out. This reduced the contig number to 30 with a 22.2 Mb genome size. Chromosome names were assigned based on alignment to Py17X. A consensus genome was generated based on the alignment of these 30 contigs to the 17X reference genome using a custom program (Supp File 1). This resulted in a final assembly of 16 contigs totaling 23.08 Mb. The apicoplast was circularized using Circlator (v.1.5.5) (58), followed by manual correction of coordinates based on Py17X alignment.

The second assembly approach utilized both Nanopore and Illumina DNA-seq reads. All Nanopore-based raw reads were first analyzed using Nanoplot (v1.33.0) (29) for quality control purposes. Nanopore reads were assembled using Flye (v2.9) (30) software which resulted in 26 contigs. The assembled contigs were scaffolded with Ragtag.py using the Py17X genome as a reference, which resulted in 17 contigs (31). The resulting assembly was polished in a multi-step approach. First, the assembly was polished using nextPolish (v1.3.1) (32). The homopolymer and long indel errors that were still present were corrected in the second step. For these, 150 base pair reads were simulated from the Py17X genome and mapped to the polished assembly in step 1 using bwa mem (59). Variants were called with freebayes (v 1.2.0) (60), and a consensus was created with bcftools (v 1.15) (61), resulting in a second round of polished assembly. To further correct errors, we mapped the Py17XNL Illumina genomic DNA to the resulting assembly, called variants, and generated a consensus. This resulted in a third-round polished assembly. Overlapping contigs were merged, and the apicoplast sequences were circularized, resulting in a genome assembly consisting of 14 nuclear chromosomes and two organellar chromosomes. The low complexity regions and tandem repeats in both assemblies were soft masked using the tantan program (62). Assembly reports were done to compare Nanopore-based and PacBio-based genomes using the Quast program. The variants between Py17X and Py17XNL genomes were obtained from the minimap2 (v2.18) (35) whole genome alignment using paftools.js. The variants were annotated, and variant effects were obtained using SnpEff (v.5.1d) (63). The assembly completeness was assessed using BUSCO (42). For this assessment, the *Plasmodium* lineage database (plasmodium_odb10), which contained 3642 sequences from 23 *Plasmodium* species, was searched to check for the presence and completeness of the single-copy marker genes.

To create gene models and assign gene names, Braker2 (v2.1.6) (39) was first used to predict genes, which was followed with the use of reciprocal blastp. Two sets of gene models were generated for this assembly. For the first set, Nanopore dRNA-Seq reads were mapped to the assembled genome with minimap2 (v2.18) (35). dRNA-Seq read alignments provided additional exon-intron evidence in Braker2-based gene model predictions. Gene names were assigned by a reciprocal blast of the predicted proteins against Py17X proteins. For the second set of gene-model predictions, both Nanopore dRNA-Seq and Illumina RNA-Seq datasets were used. Illumina RNA-seq reads were mapped to the assembled genome using Hisat2 (v.2.2.1) (64) and were merged with Nanopore dRNA-seq alignments, which were then used for Braker2 gene-model prediction. Additionally, Prokka (v 1.14.6) was used to make gene predictions in mitochondria and apicoplast. Finally, tRNAs were predicted using tRNASCAN-SE (v.2.0.9.) (65), and rRNAs were identified by a blast search of the assembled genome using Py17X rRNAs.

Data Availability

Datasets associated with this study are available using the following identifiers: SRA BioProject: PRJNA769959, Nanopore assembly accessions: CP086268-CP086283, PacBio assembly accessions:

CP115525-CP115540. All assembly files produced in this study are provided as Supplementary File 2, and will be provided to VEuPathDB/PlasmoDB for integration and community use.

Acknowledgments

We acknowledge the Penn State Genomics Core Facility - University Park for sequencing library preparations and conducting the PacBio and Illumina sequencing described in this study. We thank Tanya Renner and her laboratory at Penn State for critical discussions on Nanopore sequencing. We thank New England Biolabs for early access to high molecular weight gDNA purification kits and their insights on the optimization of this process. We also thank members of the VEuPathDB and PlasmoDB.org teams for assistance with current data files and assemblies, Akhil Vaidya (Drexel University) for discussions about the *Plasmodium* mitochondrion, Photini Sinnis on discussions about CSP and its central repeat sequence, as well as the members of the Llinás and Lindner laboratories for critical discussions of this work.

Funding

This work was supported by awards from NIAID (R01AI123341, R56AI123341) to SEL, NIGMS (T32GM125592) to MJG, and support to AS from the Huck Institutes of the Life Sciences.

Financial Disclosures

We have no financial disclosures associated with this study.

Figure Legends

Figure 1: PacBio HIFi high-quality long reads improve upon the pre-existing Py17XNL genome and outperform a hybrid assembly approach with Nanopore and Illumina sequencing. (A) QScore vs. read length distribution for a PacBio sequencing run that was used to construct the final Py17XNL_2 genome assembly is presented. Note: HiFi PacBio sequencing has a minimum QScore threshold of 20, and a maximum QScore threshold of 93. (B) A histogram is plotted to illustrate the distribution of PacBio read lengths. (C) A comparison of assembly statistics between Nanopore and PacBio sequencing runs is provided. All statistics are based on contigs of size >=500 bp. (D) The cumulative length of contigs is plotted from largest to smallest.

Figure 2: Bioinformatics workflow used for genome assembly and annotation. (Left) Genome Assembly: High-accuracy ccs reads that were generated from PacBio subreads and trimmed Nanopore reads were *de novo* assembled to create draft genomes. Contigs were selected, and chromosome names were assigned based on the *P. yoelii* 17X reference genome alignment. Further processing of the Nanopore + Illumina hybrid assembly involved implementing scaffolding and iterative polishing. (Right) Gene-model prediction: A Nanopore dRNA-seq-based gene model and a hybrid gene model combining both Nanopore dRNA-seq and Illumina RNA-seq data were generated using Braker2. The predicted genes were annotated using reciprocal BLAST against *P. yoelii* 17X proteins. Illumina RNA-seq reads were previously reported (37).

Figure 3: Expanded *Plasmodium yoelii* 17XNL gene models leveraging RNA-seq data. (A) An example gene model depicting IMC1a and its respective sequence features is provided. (B) The 3'UTR length distribution of all detected mRNAs is plotted as a histogram for chromosomal and mitochondrial genes. Transcripts encoded by the apicoplast are not polyadenylated and were not detected by Nanopore dRNA-seq. (C) The maximum, average, median, and mode of the 3' UTR lengths from all chromosomal and mitochondrial transcripts are compared to those from a *Plasmodium falciparum* dataset (13).

Figure 4: BUSCO analysis demonstrates genome assembly completeness. Of the 3,642 BUSCO groups that were searched, 3,556 single-copy BUSCOs were found to be present in the 17XNL_2 assembly resulting in a completeness score of 97.6%. The BUSCO results for Py17XNL_1 (83.9%) and Py17X (98.0%) reference genomes are also shown for comparison.

Figure 5: Differences between the *P. yoelii* 17X and 17XNL_2 assemblies. (A) The Py17XNL_2 reference genome was mapped to Py17X to determine their degree of similarity. A dot plot depicting this agreement is shown, with blue lines denoting unique alignments and orange lines depicting repeat regions. (B) A circos plot is presented with the following tracks listed from outside to inside: 1) Py17X reference genome, 2) Py17XNL_2 ccs read coverage in the natural log scale (minimum value of 0 and maximum value of 8), 3) SNPs and indels between the two genomes are shown in light green, 4) SNPs and indels in the coding sequence of genes are shown in orange. An expanded view that includes the apicoplast and mitochondria is shown separately.

Figure 6: Identification of blood stage-expressed variants between 17X and 17XNL_2. (Left) Variants of interest that are expressed in blood-stage parasites were chosen based on the presence of the variant sequence within the coding sequence, the extent to which the variant calls are supported by sequencing data, and if the gene has been named. Downselected genes are further described in Supplemental Table 4. To be considered, at least two sequencing methods needed to support the variant call, with at least 80% of the reads in agreement and a minimum of five reads at the position (three read minimum for Nanopore). (Right) IGV snapshots with representative examples of different variants found in AP2-SP (PY17XNL_1303202), RAD50 (PY17XNL_0104722), or CSP (PY17XNL_0404050) are presented top to bottom.

Figure 7: The location and potential impact on translation of variants between 17X and 17XNL_2 genome assemblies. (A) The distribution of variant locations throughout the entire Py17XNL_2 genome is shown. (B) The types of variants represented within the Py17XNL_2 genome with their respective counts are plotted. (C) The distribution of variant types within coding sequences is depicted as a bar graph.

Table 1: Summary of finalized genome assembly and gene model creation statistics. * Determined using the Quast program by alignment to the 17X reference for this study. ** Determined using a BUSCO analysis for this study. *** As reported in Carlton et al Nature 2002. N/D: Not determined.

Table 2: Prioritized list of coding sequence variants between the Py17X and Py17XNL genome.

Supplementary Figure 1: Experimental workflow for all sequencing runs performed. Four different sample/sequencing types were generated. For each, mice were infected with Py17XNL strain parasites until parasitemia reached 1-3%, at which point blood was collected, passed through a cellulose column, and saponin lysed prior to DNA or RNA recovery. For Illumina, PacBio, and Nanopore DNA samples, the NEB Monarch High Molecular Weight Blood Kit was used. For Nanopore RNA samples, a Qiagen RNeasy Kit with subsequent DNaseI treatment was used. For quality control purposes, a Nanodrop and Qubit were used to assess each biological sample. Additionally, TapeStation and Bionalyzer were used for DNA and RNA samples, respectively. The library preparation methods and sequencing devices used for each sample are also indicated. The Illumina RNA-seq data utilized in this study was previously published by our laboratories and was retrieved from the GEO depository (Accession #GSE136674) (37).

Supplementary Figure 2: Determination of gDNA fragment length by TapeStation. (A) The Qiagen Blood Amp Kit or the NEB Monarch High Molecular Weight Blood Kit were used to prepare gDNA and samples were run in parallel on an Agilent TapeStation 4150. High molecular weight gDNA from lane C1 was used for Nanopore replicate one. (B) High molecular weight gDNA used for Nanopore replicate two was run separately on the same Agilent TapeStation 4150 instrument. The hazard symbol in lane B1 indicates the sample was run outside of the manufacturer’s recommended concentration. (C) High molecular weight gDNA that was used for PacBio HiFi sequencing is shown in lane C2 on the right. All other lanes were samples from unrelated experiments.

Supplementary Figure 3: A comparison of fast basecalling and high accuracy basecalling for Nanopore DNA sequencing. (A and B) Nanopore ligation sequencing reads from replicate one were basecalled using the fast basecalling algorithm (A) or the high accuracy basecalling algorithm (B). The Qscore vs. read length distribution is depicted as a scatter plot (top), and the read length and their respective counts are plotted as a histogram (bottom). (C and D) The same comparisons as described in A and B were applied to replicate two.

Supplementary Figure 4: Bioanalyzer results demonstrate that RNA samples are of high quality. (A) Total RNA isolated for replicate one of Nanopore direct RNA sequencing was run on a Bioanalyzer for quality control purposes. The yellow hazard sign in lane B1 indicates that the markers ran outside of their standard position, leading to an edited RIN. The sample was also run at a 1:5 dilution. (B) The RNA sample used for replicate 2 of Nanopore direct RNA sequencing was run separately in the same way.

Supplementary Figure 5: A comparison of fast basecalling and high accuracy basecalling for Nanopore direct RNA sequencing. Nanopore direct RNA sequencing reads from replicate one (A and B) or two (C and D) were basecalled using the fast basecalling algorithm or the high accuracy basecalling algorithm. The Qscore vs. read length distribution is depicted as a scatter plot (top), and the read length and their respective counts are plotted as a histogram (bottom).

Supplementary Figure 6: The CSP central repeat region length varies across sequenced *P. yoelii* strains. The amino acid sequence for the central repeat region of circumsporozoite protein (CSP) is shown for *P. yoelii* 17X, *P. yoelii* 17XNL, and *P. yoelii* YM.

Supplementary Table 1: Quality control measurements from Nanodrop, Qubit, TapeStation, and Bioanalyzer.

Supplementary Table 2: Nanopore sequencing statistics for replicates 1 and 2 with the fast basecaller (LA) and the high accuracy (HA) basecaller.

Supplementary Table 3: 5' and 3' untranslated region (UTR) information by gene name using Nanopore and Illumina ("hybrid") or Nanopore-only approaches.

Supplementary Table 4: Complete list of identified CDS variants with variant sequence and sequencing support information.

Supplementary File 1: Makefile that details the bioinformatics workflow used in this study.

Supplementary File 2: All assembly files generated in this study, including genome fasta, transcript fasta, cds fasta, protein fasta, and GFF3 files.

References:

1. Organization WH. World Malaria Report. 2022.

2. De Niz M, Heussler VT. Rodent malaria models: insights into human disease and parasite biology. Curr Opin Microbiol. 2018;46:93-101.

3. Voza T, Miller JL, Kappe SH, Sinnis P. Extrahepatic exoerythrocytic forms of rodent malaria parasites at the site of inoculation: clearance after immunization, susceptibility to primaquine, and contribution to blood-stage infection. Infect Immun. 2012;80(6):2158-64.

4. Lindner SE, Swearingen KE, Shears MJ, Walker MP, Vrana EN, Hart KJ, et al. Transcriptomics and proteomics reveal two waves of translational repression during the maturation of malaria parasite sporozoites. Nat Commun. 2019;10(1):4964.

5. Swearingen KE, Lindner SE. Plasmodium Parasites Viewed through Proteomics. Trends Parasitol. 2018;34(11):945-60.

6. Munoz EE, Hart KJ, Walker MP, Kennedy MF, Shipley MM, Lindner SE. ALBA4 modulates its stage-specific interactions and specific mRNA fates during Plasmodium yoelii growth and transmission. Mol Microbiol. 2017;106(2):266-84.

7. Li J, Cai B, Qi Y, Zhao W, Liu J, Xu R, et al. UTR introns, antisense RNA and differentially spliced transcripts between Plasmodium yoelii subspecies. Malar J. 2016;15:30.

8. Lindner SE, Mikolajczak SA, Vaughan AM, Moon W, Joyce BR, Sullivan WJ, Jr., et al. Perturbations of Plasmodium Puf2 expression and RNA-seq of Puf2-deficient sporozoites reveal a critical role in maintaining RNA homeostasis and parasite transmissibility. Cell Microbiol. 2013;15(7):1266-83.

9. Ogun SA, Tewari R, Otto TD, Howell SA, Knuepfer E, Cunningham DA, et al. Targeted disruption of py235ebp-1: invasion of erythrocytes by Plasmodium yoelii using an alternative Py235 erythrocyte binding protein. PLoS Pathog. 2011;7(2):e1001288.

10. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, et al. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature. 2002;419(6906):512-9.

11. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002;419(6906):498-511.

12. Yang M, Shang X, Zhou Y, Wang C, Wei G, Tang J, et al. Full-Length Transcriptome Analysis of Plasmodium falciparum by Single-Molecule Long-Read Sequencing. Front Cell Infect Microbiol. 2021;11:631545.

13. Chappell L, Ross P, Orchard L, Russell TJ, Otto TD, Berriman M, et al. Refining the transcriptome of the human malaria parasite Plasmodium falciparum using amplification-free RNA-seq. BMC Genomics. 2020;21(1):395.

14. Fougere A, Jackson AP, Bechtsi DP, Braks JA, Annoura T, Fonager J, et al. Variant Exported Blood-Stage Proteins Encoded by Plasmodium Multigene Families Are Expressed in Liver Stages Where They Are Exported into the Parasitophorous Vacuole. PLoS Pathog. 2016;12(11):e1005917.

15. Otto TD, Bohme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, et al. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. 2014;12:86.

16. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, et al. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science. 2005;307(5706):82-6.

17. Zhang C, Oguz C, Huse S, Xia L, Wu J, Peng YC, et al. Genome sequence, transcriptome, and annotation of rodent malaria parasite Plasmodium yoelii nigeriensis N67. BMC Genomics. 2021;22(1):303.

18. Wang J, Chen K, Yang J, Zhang S, Li Y, Liu G, et al. Comparative genomic analysis of Babesia duncani responsible for human babesiosis. BMC Biol. 2022;20(1):153.

19. Menon VK, Okhuysen PC, Chappell CL, Mahmoud M, Mahmoud M, Meng Q, et al. Fully resolved assembly of Cryptosporidium parvum. Gigascience. 2022;11.

20. Baptista RP, Li Y, Sateriale A, Sanders MJ, Brooks KL, Tracey A, et al. Long-read assembly and comparative evidence-based reanalysis of Cryptosporidium genome sequences reveal expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions. Genome Res. 2022;32(1):203-13.

21. Toro-Moreno M, Sylvester K, Srivastava T, Posfai D, Derbyshire ER. RNA-Seq Analysis Illuminates the Early Stages of Plasmodium Liver Infection. mBio. 2020;11(1).

22. Caldelari R, Dogga S, Schmid MW, Franke-Fayard B, Janse CJ, Soldati-Favre D, et al. Transcriptome analysis of Plasmodium berghei during exo-erythrocytic development. Malar J. 2019;18(1):330.

23. Kent RS, Modrzynska KK, Cameron R, Philip N, Billker O, Waters AP. Inducible developmental reprogramming redefines commitment to sexual development in the malaria parasite Plasmodium berghei. Nat Microbiol. 2018;3(11):1206-13.

24. Yeoh LM, Goodman CD, Mollard V, McFadden GI, Ralph SA. Comparative transcriptomics of female and male gametocytes in Plasmodium berghei and the evolution of sex in alveolates. BMC Genomics. 2017;18(1):734.

25. Mancio-Silva L, Slavic K, Grilo Ruivo MT, Grosso AR, Modrzynska KK, Vera IM, et al. Nutrient sensing modulates malaria parasite virulence. Nature. 2017;547(7662):213-6.

26. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348-65.

27. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30.

28. Godin MJL, S.E. NEB Monarch® HMW DNA Extraction Kit improves sample preparation for Oxford Nanopore Technologies sequencing of malaria parasites2021. Available from: https://www.neb.com/-/media/nebus/files/application-notes/appnote_monarch_hmw_dna_improves_sample_prep_for-ont_sequencing_of_malaria.pdf?rev=9c525dc516834ba684bf168dc13a164b

29. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666-9.

30. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540-6.

31. Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB, et al. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. bioRxiv. 2021:2021.11.18.469135.

32. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36(7):2253-5.

33. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.

34. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722-36.

35. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.

36. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37(Database issue):D539-43.

37. Hart KJ, Power BJ, Rios KT, Sebastian A, Lindner SE. The Plasmodium NOT1-G paralogue is an essential regulator of sexual stage maturation and parasite transmission. PLoS Biol. 2021;19(10):e3001434.

38. Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, et al. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification. Elife. 2020;9.

39. Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3(1):lqaa108.

40. Gish W, States DJ. Identification of protein coding regions by database similarity search. Nat Genet. 1993;3(3):266-72.

41. Gebauer F, Preiss T, Hentze MW. From cis-regulatory elements to complex RNPs and back. Cold Spring Harb Perspect Biol. 2012;4(7):a012245.

42. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.

43. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I. Transcription factor AP2-Sp and its target genes in malarial sporozoites. Mol Microbiol. 2010;75(4):854-63.

44. Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res. 2005;33(13):3994-4006.

45. Martins RM, Macpherson CR, Claes A, Scheidig-Benatar C, Sakamoto H, Yam XY, et al. An ApiAP2 member regulates expression of clonally variant genes of the human malaria parasite Plasmodium falciparum. Sci Rep. 2017;7(1):14042.

46. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, et al. Specific DNA-binding by apicomplexan AP2 transcription factors. Proc Natl Acad Sci U S A. 2008;105(24):8393-8.

47. Russell TJ, De Silva EK, Crowley VM, Shaw-Saliba K, Dube N, Josling G, et al. Inhibitors of ApiAP2 protein DNA binding exhibit multistage activity against Plasmodium parasites. PLoS Pathog. 2022;18(10):e1010887.

48. Zhang M, Mishra S, Sakthivel R, Rojas M, Ranjan R, Sullivan WJ, Jr., et al. PK4, a eukaryotic initiation factor 2alpha(eIF2alpha) kinase, is essential for the development of the erythrocytic cycle of Plasmodium. Proc Natl Acad Sci U S A. 2012;109(10):3956-61.

49. McDowall J, Hunter S. InterPro protein classification. Methods Mol Biol. 2011;694:37-47.

50. Zhang M, Fennell C, Ranford-Cartwright L, Sakthivel R, Gueirard P, Meister S, et al. The Plasmodium eukaryotic initiation factor-2alpha kinase IK2 controls the latency of sporozoites in the mosquito salivary glands. J Exp Med. 2010;207(7):1465-74.

51. Mohrle JJ, Zhao Y, Wernli B, Franklin RM, Kappes B. Molecular cloning, characterization and localization of PfPK4, an eIF-2alpha kinase-related enzyme from the malarial parasite Plasmodium falciparum. Biochem J. 1997;328 (Pt 2)(Pt 2):677-87.

52. Balaban AE, Kanatani S, Mitra J, Gregory J, Vartak N, Sinnis-Bourozikas A, et al. The repeat region of the circumsporozoite protein is an elastic linear spring with a functional role in <em>Plasmodium</em> sporozoite motility. bioRxiv. 2021:2021.05.12.443759.

53. Bryant JM, Baumgarten S, Lorthiois A, Scheidig-Benatar C, Claes A, Scherf A. De Novo Genome Assembly of a Plasmodium falciparum NF54 Clone Using Single-Molecule Real-Time Sequencing. Genome Announc. 2018;6(5).

54. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53.

55. Hoshizaki J, Adjalley SH, Thathy V, Judge K, Berriman M, Reid AJ, et al. A manually curated annotation characterises genomic features of P. falciparum lncRNAs. BMC Genomics. 2022;23(1):780.

56. Shaw PJ, Kaewprommal P, Wongsombat C, Ngampiw C, Taechalertpaisarn T, Kamchonwongpaisan S, et al. Transcriptomic complexity of the human malaria parasite Plasmodium falciparum revealed by long-read sequencing. PLoS One. 2022;17(11):e0276956.

57. Lee VV, Judd LM, Jex AR, Holt KE, Tonkin CJ, Ralph SA. Direct Nanopore Sequencing of mRNA Reveals Landscape of Transcript Isoforms in Apicomplexan Parasites. mSystems. 2021;6(2).

58. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015;16:294.

59. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

60. Said Mohammed K, Kibinge N, Prins P, Agoti CN, Cotten M, Nokes DJ, et al. Evaluating the performance of tools used to call minority variants from whole genome short-read data. Wellcome Open Res. 2018;3:21.

61. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).

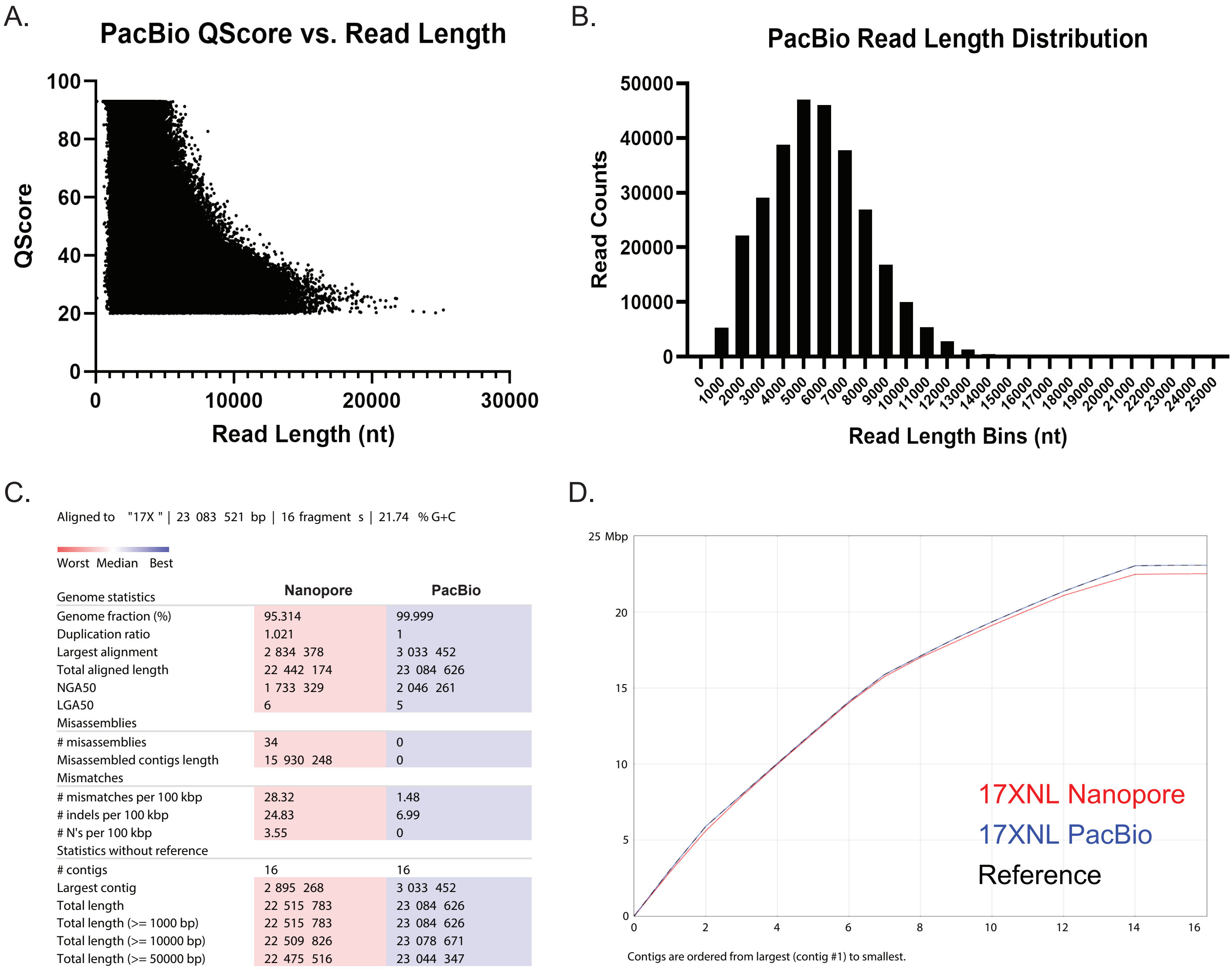
62. Frith MC. A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. 2011;39(4):e23.

63. Cingolani P. Variant Annotation and Functional Prediction: SnpEff. Methods Mol Biol. 2022;2493:289-314.

64. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-15.

65. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955-64.

Godin and Sebastian *et al*. Figure 1



Aligned to "17X" | 23 083 521 bp | 16 fragments | 21.74 % G+C

Genome statistics	Nanopore	PacBio
Genome fraction (%)	95.314	99.999
Duplication ratio	1.021	1
Largest alignment	2 834 378	3 033 452
Total aligned length	22 442 174	23 084 626
NGA50	1 733 329	2 046 261
LGA50	6	5
Misassemblies		
# misassemblies	34	0
Misassembled contigs length	15 930 248	0
Mismatches		
# mismatches per 100 kbp	28.32	1.48
# indels per 100 kbp	24.83	6.99
# N's per 100 kbp	3.55	0
Statistics without reference		
# contigs	16	16
Largest contig	2 895 268	3 033 452
Total length	22 515 783	23 084 626
Total length (>= 1000 bp)	22 515 783	23 084 626
Total length (>= 10000 bp)	22 509 826	23 078 671
Total length (>= 50000 bp)	22 475 516	23 044 347

Figure 1: PacBio HIFi high-quality long reads improve upon the pre-existing Py17XNL genome and outperform a hybrid assembly approach with Nanopore and Illumina sequencing. (A) QScore vs. read length distribution for a PacBio sequencing run that was used to construct the final Py17XNL_2 genome assembly is presented. Note: HiFi PacBio sequencing has a minimum QScore threshold of 20, and a maximum QScore threshold of 93. (B) A histogram is plotted to illustrate the distribution of PacBio read lengths. (C) A comparison of assembly statistics between Nanopore and PacBio sequencing runs is provided. All statistics are based on contigs of size >=500 bp. (D) The cumulative length of contigs is plotted from largest to smallest.

Godin and Sebastian *et al*. Figure 2

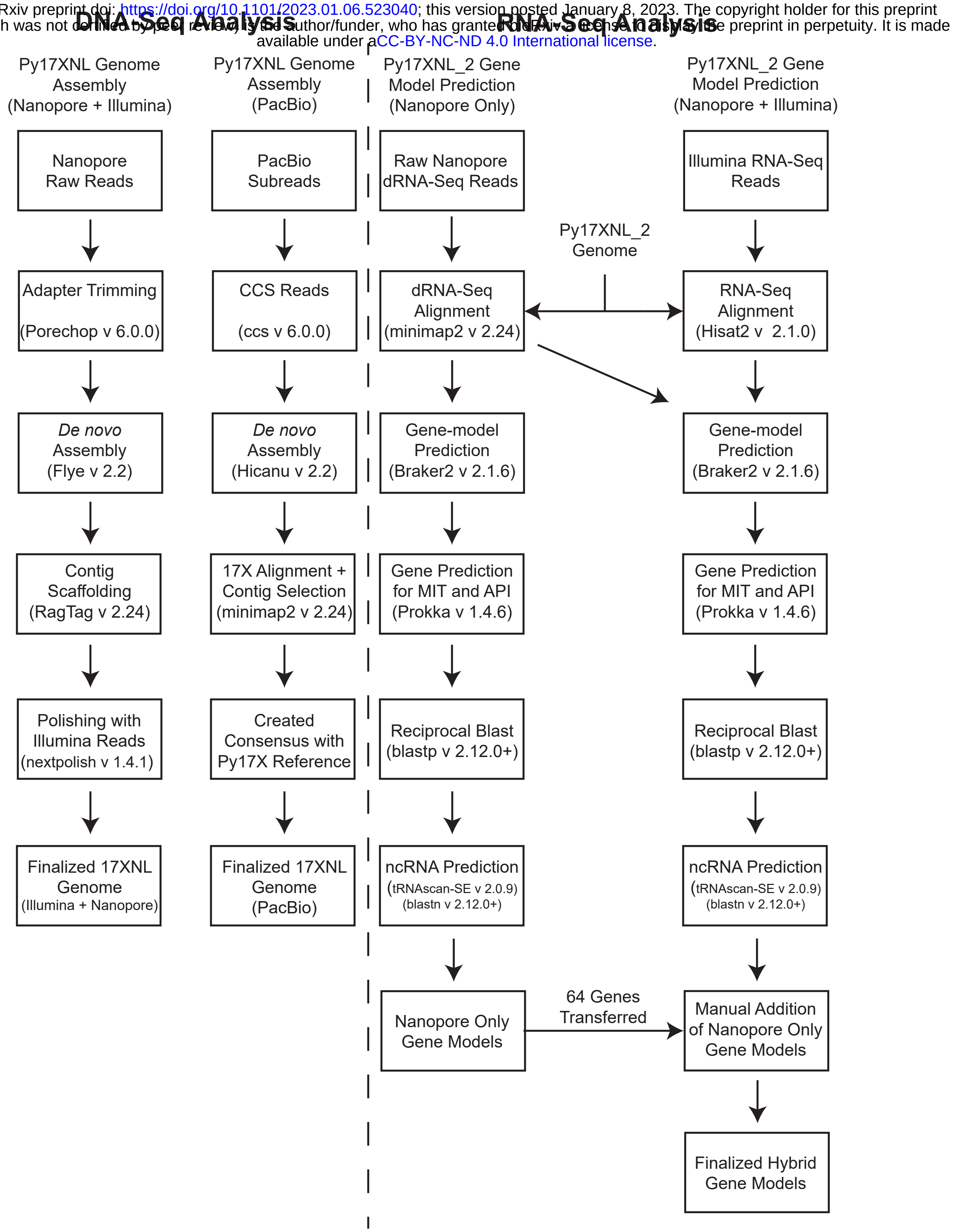
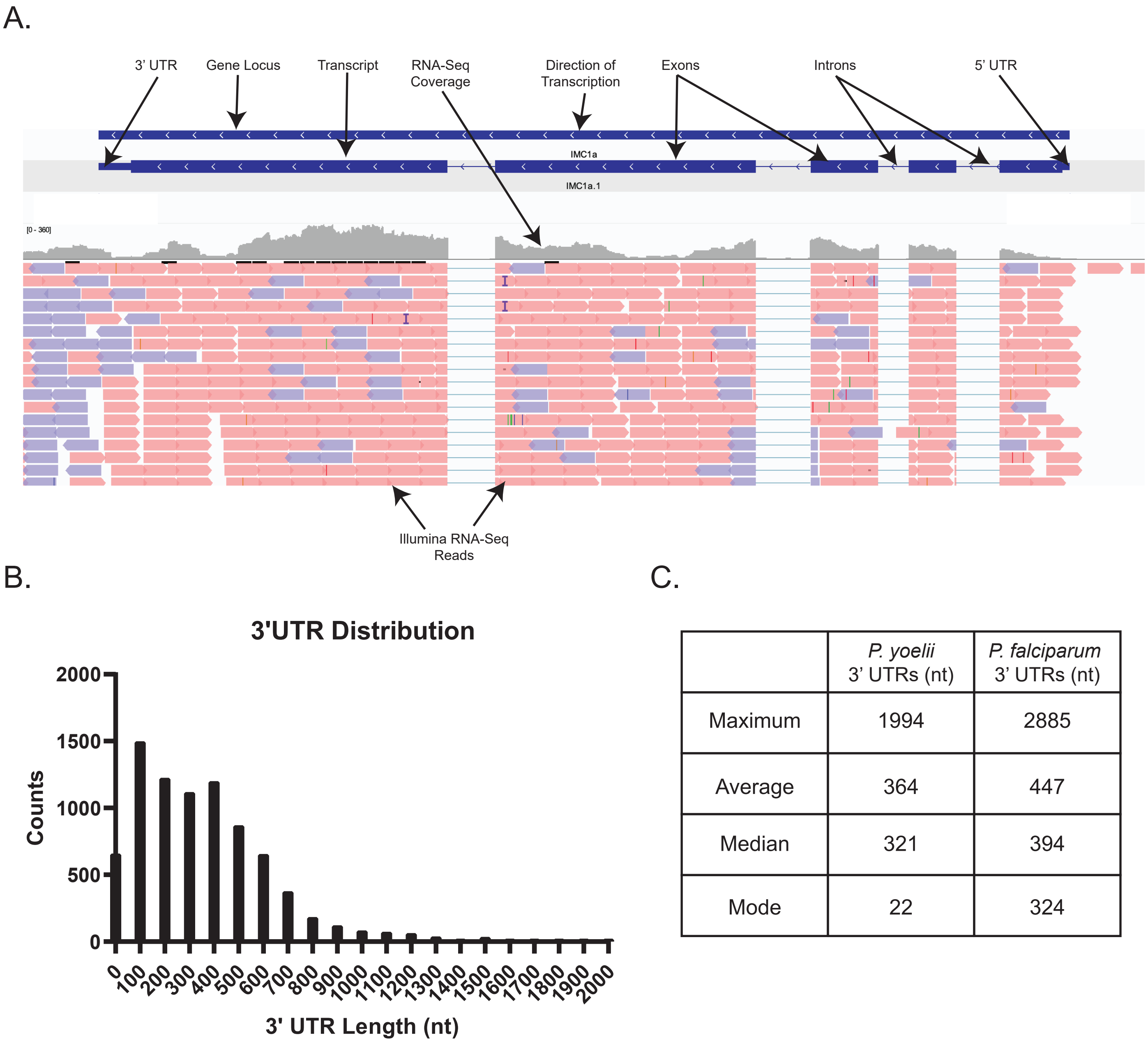

DNA-Seq Analysis
RNA-Seq Analysis
Py17XNL Genome Assembly (Nanopore + Illumina)
Py17XNL Genome Assembly (PacBio)
Py17XNL_2 Gene Model Prediction (Nanopore Only)
Py17XNL_2 Gene Model Prediction (Nanopore + Illumina)
Nanopore Raw Reads
PacBio Subreads
Raw Nanopore dRNA-Seq Reads
Illumina RNA-Seq Reads
Py17XNL_2 Genome
Adapter Trimming (Porechop v 6.0.0)
CCS Reads (ccs v 6.0.0)
dRNA-Seq Alignment (minimap2 v 2.24)
RNA-Seq Alignment (Hisat2 v 2.1.0)
De novo Assembly (Flye v 2.2)
De novo Assembly (Hicanu v 2.2)
Gene-model Prediction (Braker2 v 2.1.6)
Gene-model Prediction (Braker2 v 2.1.6)
Contig Scaffolding (RagTag v 2.24)
17X Alignment + Contig Selection (minimap2 v 2.24)
Gene Prediction for MIT and API (Prokka v 1.4.6)
Gene Prediction for MIT and API (Prokka v 1.4.6)
Polishing with Illumina Reads (nextpolish v 1.4.1)
Created Consensus with Py17X Reference
Reciprocal Blast (blastp v 2.12.0+)
Reciprocal Blast (blastp v 2.12.0+)
Finalized 17XNL Genome (Illumina + Nanopore)
Finalized 17XNL Genome (PacBio)
ncRNA Prediction (tRNAscan-SE v 2.0.9) (blastn v 2.12.0+)
ncRNA Prediction (tRNAscan-SE v 2.0.9) (blastn v 2.12.0+)
Nanopore Only Gene Models
64 Genes Transferred
Manual Addition of Nanopore Only Gene Models
Finalized Hybrid Gene Models

Figure 2: Bioinformatics workflow used for genome assembly and annotation.
(Left) Genome Assembly: High-accuracy ccs reads that were generated from PacBio subreads and trimmed Nanopore reads were de novo assembled to create draft genomes. Contigs were selected, and chromosome names were assigned based on the *P. yoelii* 17X reference genome alignment. Further processing of the Nanopore + Illumina hybrid assembly involved implementing scaffolding and iterative polishing. (Right) Gene-model prediction: A Nanopore dRNA-seq-based gene model and a hybrid gene model combining both Nanopore dRNA-seq and Illumina RNA-seq data were generated using Braker2. The predicted genes were annotated using reciprocal BLAST against *P. yoelii* 17X proteins. Illumina RNA-seq reads were previously reported (37).

Godin and Sebastian *et al*. Figure 3



	P. yoelii 3' UTRs (nt)	*P. falciparum* 3' UTRs (nt)
Maximum	1994	2885
Average	364	447
Median	321	394
Mode	22	324

Figure 3: Expanded *Plasmodium yoelii* 17XNL gene models leveraging RNA-seq data. (A) An example gene model depicting IMC1a and its respective sequence features is provided. (B) The 3'UTR length distribution of all detected mRNAs is plotted as a histogram for chromosomal and mitochondrial genes. Transcripts encoded by the apicoplast are not polyadenylated and were not detected by Nanopore dRNA-seq. (C) The maximum, average, median, and mode of the 3' UTR lengths from all chromosomal and mitochondrial transcripts are compared to those from a *Plasmodium falciparum* dataset (13).

Godin and Sebastian *et al.* Figure 4

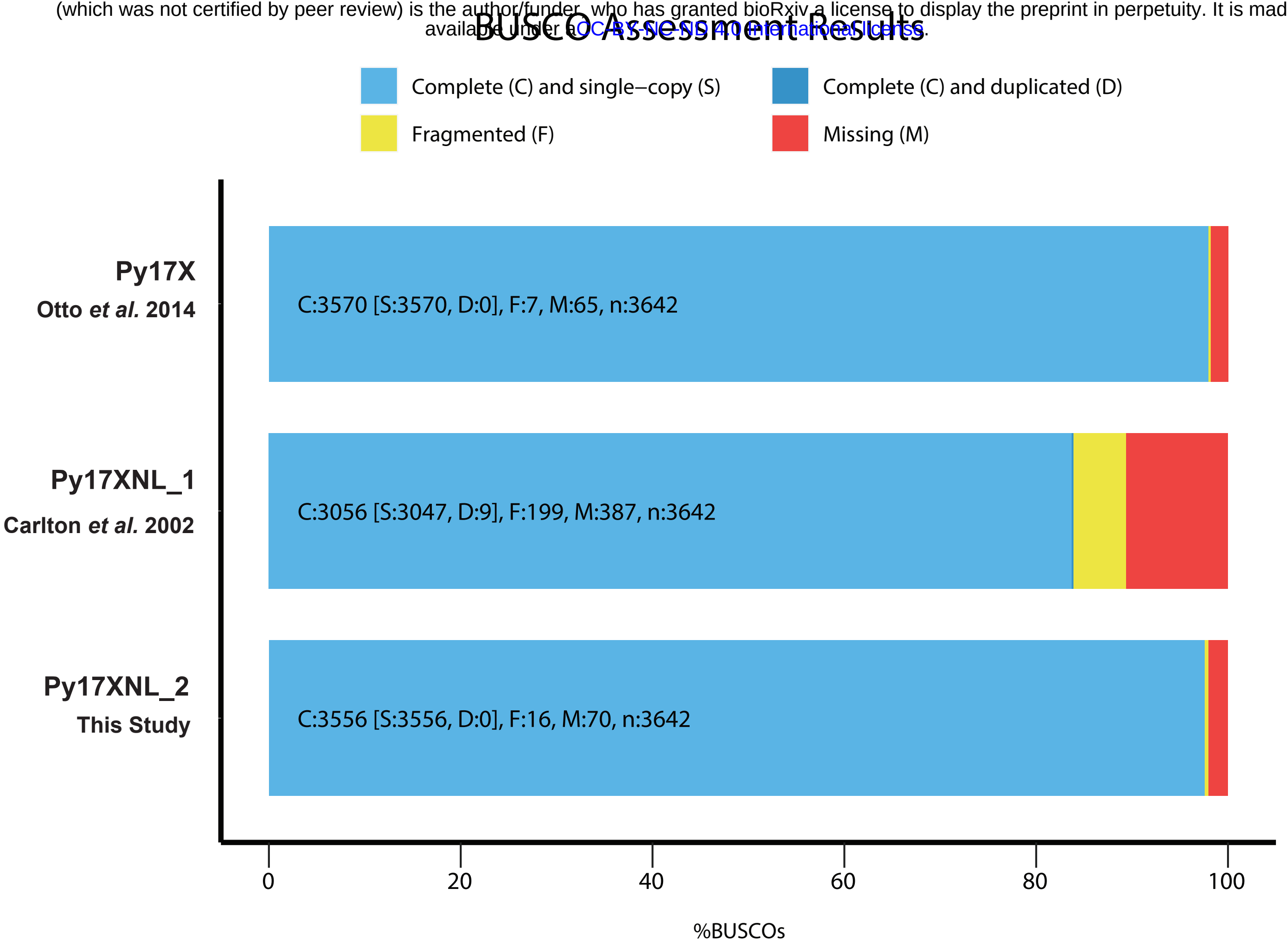


Figure 4: BUSCO analysis demonstrates genome assembly completeness.
Of the 3,642 BUSCO groups that were searched, 3,556 single-copy BUSCOs were found to be present in the 17XNL_2 assembly resulting in a completeness score of 97.6%. The BUSCO results for Py17XNL_1 (83.9%) and Py17X (98.0%) reference genomes are also shown for comparison.

Godin and Sebastian *et al*. Figure 5

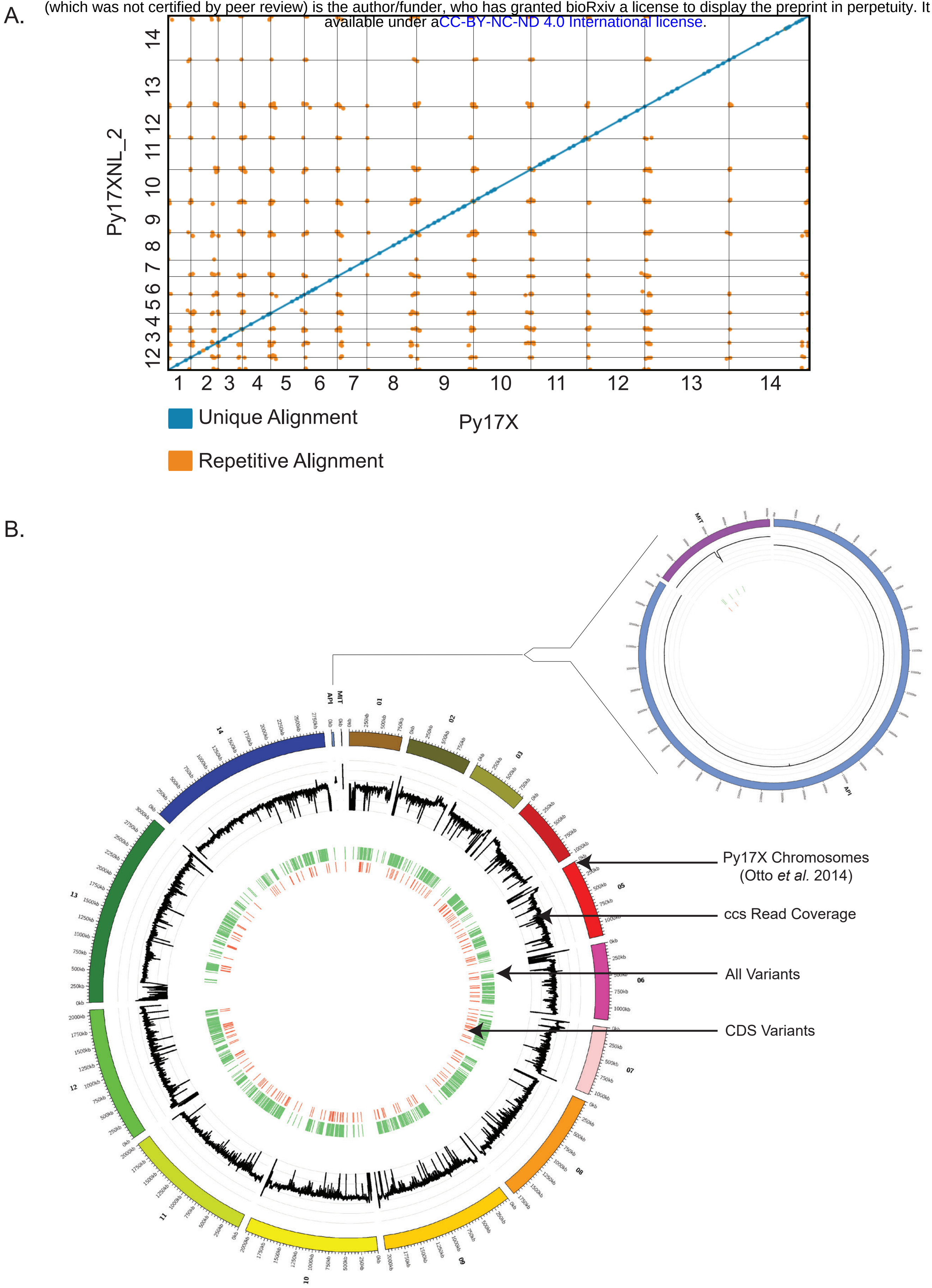


Figure 5: Differences between the *P. yoelii* 17X and 17XNL_2 assemblies. (A) The Py17XNL_2 reference genome was mapped to Py17X to determine their degree of similarity. A dot plot depicting this agreement is shown, with blue lines denoting unique alignments and orange lines depicting repeat regions. (B) A circos plot is presented with the following tracks listed from outside to inside: 1) Py17X reference genome, 2) Py17XNL_2 ccs read coverage in the natural log scale (minimum value of 0 and maximum value of 8), 3) SNPs and indels between the two genomes are shown in light green, 4) SNPs and indels in the coding sequence of genes are shown in orange. An expanded view that includes the apicoplast and mitochondria is shown separately.

Godin and Sebastian *et al*. Figure 6

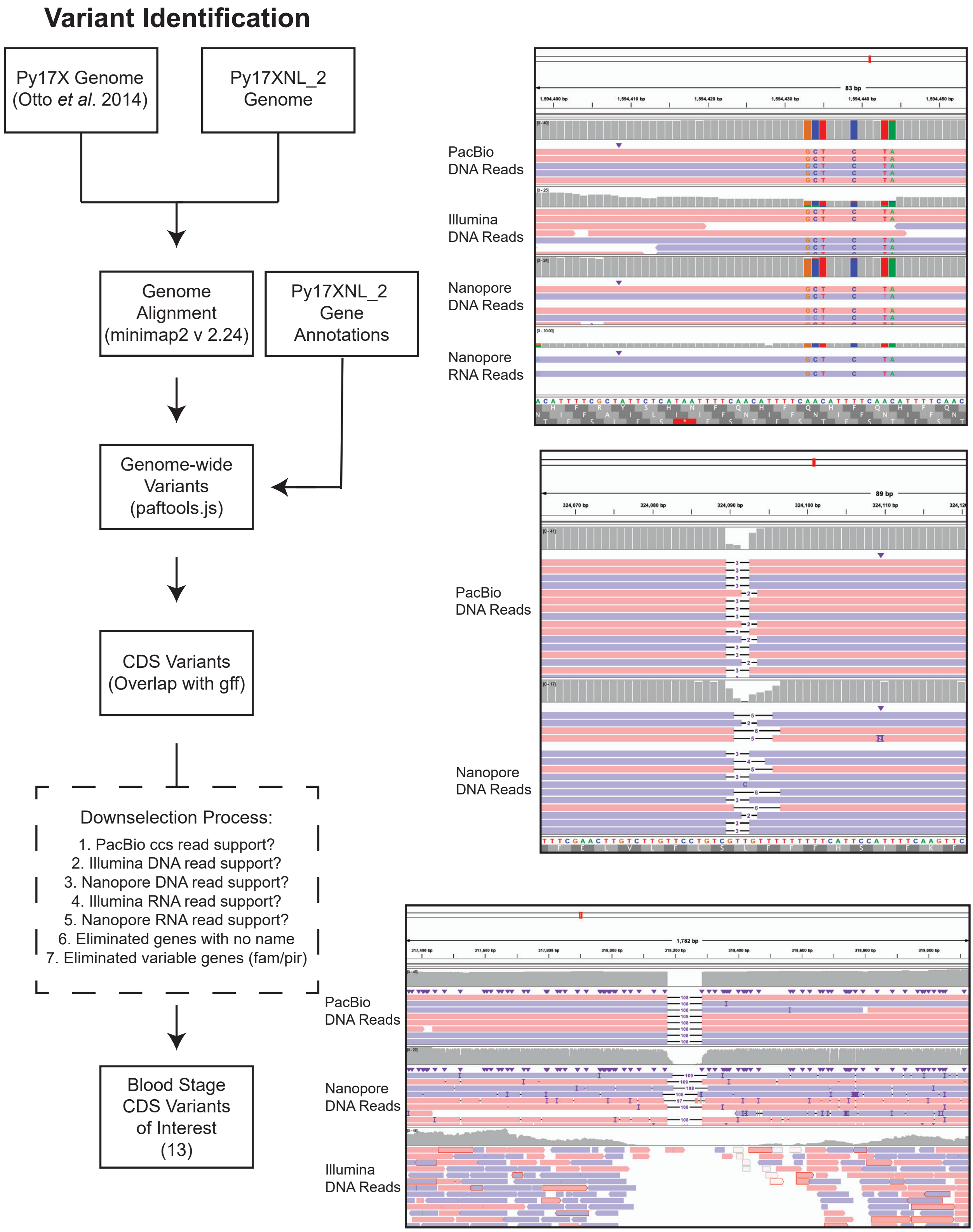

Variant Identification
Py17X Genome (Otto et al. 2014)
Py17XNL_2 Genome
Genome Alignment (minimap2 v 2.24)
Py17XNL_2 Gene Annotations
Genome-wide Variants (paftools.js)
CDS Variants (Overlap with gff)
Downselection Process:
1. PacBio ccs read support?
2. Illumina DNA read support?
3. Nanopore DNA read support?
4. Illumina RNA read support?
5. Nanopore RNA read support?
6. Eliminated genes with no name
7. Eliminated variable genes (fam/pir)
Blood Stage CDS Variants of Interest (13)
PacBio DNA Reads
Illumina DNA Reads
Nanopore DNA Reads
Nanopore RNA Reads
83 bp
PacBio DNA Reads
Nanopore DNA Reads
89 bp
PacBio DNA Reads
Nanopore DNA Reads
Illumina DNA Reads
1,782 bp

Figure 6: Identification of blood stage-expressed variants between 17X and 17XNL_2. (Left) Variants of interest that are expressed in blood-stage parasites were chosen based on the presence of the variant sequence within the coding sequence, the extent to which the variant calls are supported by sequencing data, and if the gene has been named. Downselected genes are further described in Supplemental Table 4. To be considered, at least two sequencing methods needed to support the variant call, with at least 80% of the reads in agreement and a minimum of five reads at the position (three read minimum for Nanopore). (Right) IGV snapshots with representative examples of different variants found in AP2-SP (PY17XNL_1303202), RAD50 (PY17XNL_0104722), or CSP (PY17XNL_0404050) are presented top to bottom.

Godin and Sebastian *et al*. Figure 7

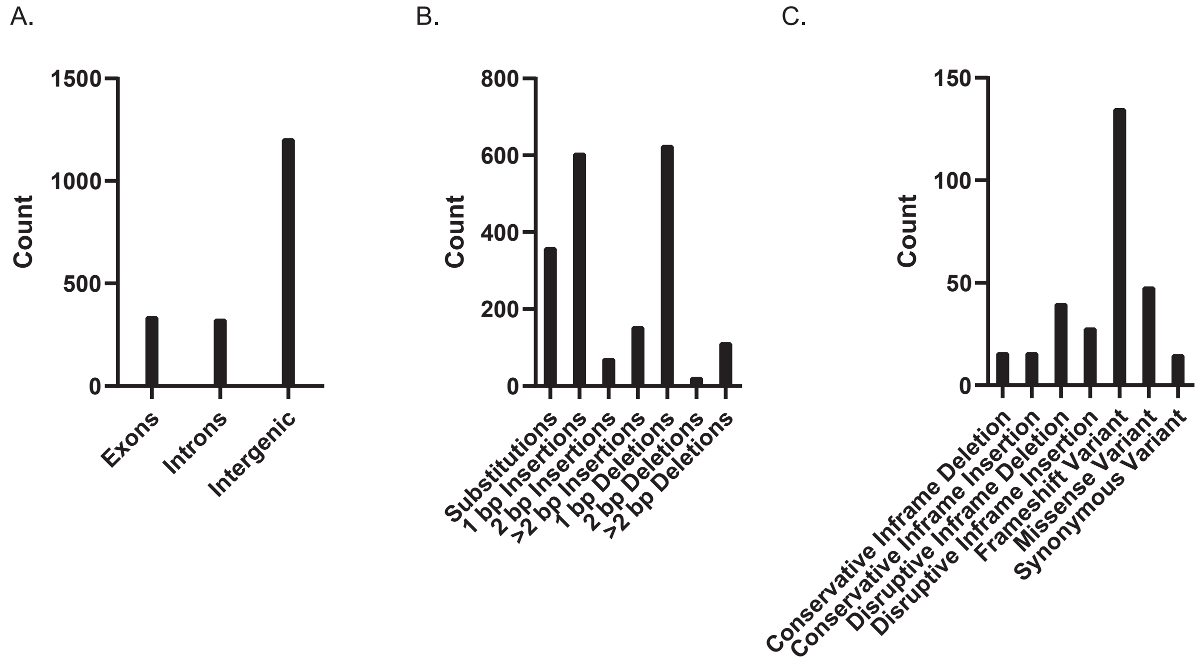


Figure 7: The location and potential impact on translation of variants between 17X and 17XNL_2 genome assemblies. (A) The distribution of variant locations throughout the entire Py17XNL_2 genome is shown. (B) The types of variants represented within the Py17XNL_2 genome with their respective counts are plotted. (C) The distribution of variant types within coding sequences is depicted as a bar graph.

Table 1.

	17X (Otto *et al.* 2014)	17XNL_1 (Carlton *et al.* 2002)	17XNL_2 (This Study)
Quast summary			
Number of contigs	16	5617 (5687***)	16
Assembled genome size	23.08Mb	23.1 Mb	23.08 Mb
Genome Fraction (%)	-	88.68% *	99.90%
Largest Alignment	-	51.5 Kb	3,033,452
NGA50	-	7,668 *	2,046,261
LGA50	-	851 *	5
Mismatches per 100 kb	-	72.72 *	1.48
Indels per 100 kb	-	44.07 *	6.99
N's per 100 kb	-	50 *	0
BUSCO summary			
BUSCO genome completeness	98% **	83.9% **	97.60%
BUSCO protein completeness	99% **	64.7% **	90.00%
Annotation summary			
Number of genes	6,263	7774 (5878***)	6,086
Number of mRNAs	6,041	7,724	7,052
Number of tRNAs	79	50 (39***)	66
Number of rRNAs	40	0 (7***)	40
Variant summary			
Variants with respect to 17X	-	N/D	1,955
Number of substitutions	-	N/D	360
Number of insertions	-	N/D	833
Number of deletions	-	N/D	762
Number of variants in CDS	-	N/D	334

* Determined using the Quast program for this study.

** Determined using a BUSCO analysis for this study.

*** As reported in Carlton et al Nature 2002

N/D Not Determined

Table 2

Py17X Gene ID	Py17XNL Gene ID	Gene Name	Chromosome Position	Py17X DNA	Py17XNL DNA	Mutation Type	AA Change	Protein Alignment
PY17X_0811400	Py17XNL_0801678	proteasome subunit alpha type-3.1	529296	G	A	Missense Variant	Val24Ile	Py17X MAGLSAGYDLSVSTFSPDGRLYQVEYIYKAINNNNTSISLECKDGVISCSINTSLEKNKMIKKNSYNRIYYV 17XNL MAGLSAGYDLSVSTFSPDGRLYQIEYIYKAINNNNTSISLECKDGVISCSINTSLEKNKMIKKNSYNRIYYV Cons ***********************:**
PY17X_0316900	Py17XNL_0303734	Plasmodium exported protein	625346	G	A	Missense Variant	Gly1324Ser	Py17X PEQKENGDIGEASNNAAELKENMNDLLKDTIEISKESIKEHDAQSIMFTRKFIKHVSGYDIQKAKDHPTDED 17XNL PEQKENGDIGEASNNAAELKENMNDLLKDTIEISKESIKEHDAQSIMFTRKFIKHVSSYDIQKAKDHPTDED Cons ***.**************
PY17X_1022000	Py17XNL_1002268	PP7	964836	C	A	Missense Variant	Leu342Phe	Py17X FAFKLSNYDSVIINRGNHECSYMNEVYGFHNEVLSKYDESVFDIFQEIFELLSLSVNIQNQIFVVHGGLSRY 17XNL FAFKLSNYDSVIINRGNHECSYMNEVYGFHNEVLSKYDESVFDIFQEIFELLSFSVNIQNQIFVVHGGLSRY Cons ***:******************
PY17X_1419800	Py17XNL_1401046	ACDC domain-containing protein	839701	G	T	Missense Variant	Arg1562Ile	Py17X VNENFTAELNGVQMYNGNEKKKKKKNYSLSINKNNGNIKDNENTNEILLRYENEVYAPNNDVEKNLIEDNNI 17XNL VNENFTAELNGVQMYNGNEKKKKKKNYSLSINKNNGNIKDNENTNEILLIYENEVYAPNNDVEKNLIEDNNI Cons ***:**********************
PY17X_0106400	Py17XNL_0104725	RNA-binding protein	339741	AGATAGG	A	Disruptive Inframe Deletion	Asp578_Arg579del	Py17X RDRDRDRDRDRDRDRDRDRDRDRR 17XNL RDRDR--DRDRDRDRDRDRDRR Cons ***** *****************
PY17X_1206500	Py17XNL_1204935	UTP25	344114	GGAAAATGGGAAT	G	Disruptive Inframe Deletion	Gly163_Asn166del	Py17X ENGNENGNENGNENGNENGNENDKNGNDKNGNDKNGNDKNEASSFQSKDEIYMNILINNIKSQNEDFLNVKE 17XNL ENGNENGNENGNENGNEN----DKNGNDKNGNDKNGNDKNEASSFQSKDEIYMNILINNIKSQNEDFLNVKE Cons ****************** **
PY17X_1110800	Py17XNL_1105517	KH domain-containing protein	539935	G	GTTA	Disruptive Inframe Insertion	Asn1809dup	Py17X NNNVGRDNIIRKENKGIMMHDDKDKFSKGGNNRYFGDKTNNFNNKN-NNNNNNNNNNNNNNNNNNAKNNYLSKDSMI 17XNL NNNVGRDNIIRKENKGIMMHDDKDKFSKGGNNRYFGDKTNNFNNKNNNNNNNNNNNNNNNNNNNAKNNYLSKDSMI Cons ** ******************************
PY17X_1334500	Py17XNL_1303202	AP2-SP	1594433	AAC, T, AC	GCT, C, TA	Synonymous and Missense Variant	Val133Tyr, Val136Ser	Py17X QINYNISNDIMNTVPSTNCDVTHDSVSSVPNNAFENVENVKNVENVENVKNVENVENVENVENVENVENYEN 17XNL QINYNISNDIMNTVPSTNCDVTHDSVSSVPNNAFENVENVKNVENVENVKNVENVENVENYENSENVENYEN Cons ** ** ********
PY17X_1128400	Py17XNL_1105678	PK4	1187719	T	TGAA	Conservative Inframe Insertion	Glu266dup	Py17X FYNSYNYCNNNNSKRDEKIEKNIVEKNIENKYNIKEYDKTNKSILFPIE-EFKKIIQIENNIERNYIVPKES 17XNL FYNSYNYCNNNNSKRDEKIEKNIVEKNIENKYNIKEYDKTNKSILFPIEEEFKKIIQIENNIERNYIVPKES Cons *** **********************
PY17X_1451200	Py17XNL_1401341	BDP5	2018932	A	AAATATAAAC	Disruptive Inframe Insertion	Asn320_Asn321insAsnIleAsn	Py17X NKIRSKNEINNSPNTDKVEKNIN---NINNINNINNNTNNNNVHEYVPNNLNDEFIEEKKLDKNKFNEYKNN 17XNL NKIRSKNEINNSPNTDKVEKNINNINNINNINNNTNNNNVHEYVPNNLNDEFIEEKKLDKNKFNEYKNN Cons *********************** **
PY17X_0942100	Py17XNL_0900429	PAIP1	1662686	A	AAAT	Disruptive Inframe Insertion	Asn1941dup	Py17X NVNKNKEIGKDEIQINSQINNLDDNAKGKKSNIFNQAKSSYKYPAEEGENNSNTNTSTEN-NNNNNNNNNNKT 17XNL NVNKNKEIGKDEIQINSQINNLDDNAKGKKSNIFNQAKSSYKYPAEEGENNSNTNTSTENNNNNNNNNNNNKT Cons ** ************
PY17X_0106100	Py17XNL_0104722	RAD50	324089	CGTT	C	Disruptive Inframe Deletion	Gln834del	Py17X ENITNCVNKNEDILSDNLIKLESKKRVTAHFEELENGMKKKQRQEQDKFETVQKMKIEKIEKISKIEKINKI 17XNL ENITNCVNKNEDILSDNLIKLESKKRVTAHFEELENGMKKK-RQEQDKFETVQKMKIEKIEKISKIEKINKI Cons *** ******************************
PY17X_1001900	Py17XNL_1204888	erythrocyte membrane antigen 1	174565	TAAATGA	T	Conservative Inframe Deletion	Asn287_Glu288del	Py17X SYLNNGENAEDQELDDEVASCFADGENVNDKELDEVISYLANGENVNVNVNENVNENVNENVNENENENENE 17XNL SYLNNGENAEDQELDDEVASCFADGENVNDKELDEVISYLANGENVNVNVNENVNENVNENVNE--NENENE Cons ** ******