From contigs to chromosomes: automatic Improvement of Long Read Assemblies (ILRA)

Authors:

José L. Ruiz[1], Susanne Reimering[2], Mandy Sanders[3], Juan David Escobar-Prieto[4], Nicolas M. B. Brancucci[5,6,7], Diego F. Echeverry[4,8], Abdirahman I. Abdi[9], Matthias Marti[5], Elena Gómez-Díaz[1] and Thomas D. Otto[5*]

[1] Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones Científicas, 18016, Granada, Spain
[2] Department for Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Braunschweig, Germany
[3] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[4] Centro Internacional de Entrenamiento e Investigaciones Médicas (CIDEIM), Cali, Colombia
[5] Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, MVLS, University of Glasgow, Glasgow, UK
[6] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, 4051 Basel, Switzerland.
[7] University of Basel, 4001 Basel, Switzerland.
[8] Departamento de Microbiología, Facultad de Salud, Universidad del Valle, Cali, Colombia
[9] KEMRI-Wellcome Trust Research Programme, CGMRC, Kilifi, Kenya

* Correspondence to: Thomas D. Otto (Thomasdan.otto@glasgow.ac.uk)

ABSTRACT

Recent advances in long read technologies not only enable large consortia to aim to sequence all eukaryotes on Earth, but they also allow many laboratories to sequence their species of interest. Although there is a promise to obtain “perfect genomes” with long read technologies, the number of contigs often exceeds the number of chromosomes significantly, containing many insertion and deletion errors around homopolymer tracks. To overcome these issues, we implemented ILRA to correct long reads-based assemblies, a pipeline that orders, names, merges, and circularizes contigs, filters erroneous small contigs and contamination, and corrects homopolymer errors with Illumina reads. We successfully tested our approach to assemble the genomes of four novel *Plasmodium falciparum* samples, and on existing assemblies of *Trypanosoma brucei* and *Leptosphaeria* spp. We found that correcting homopolymer tracks reduced the number of genes incorrectly annotated as pseudogenes, but an iterative correction seems to be needed to reduce high numbers of homopolymer errors. In summary, we described and compared the performance of a new tool, which

improves the quality of long read assemblies. It can be used to correct genomes of a size of up to 300 Mb.
Availability: The tool is available at GitHub: https://github.com/ThomasDOtto/ILRA.

Keywords: *de novo* assembly, automatic finishing, software, sequence improvement

INTRODUCTION

Next Generation Sequencing techniques have undergone impressive development over the last few years, achieving unparalleled resolution and performance (Marx, 2021). The extended read lengths from Pacific Bioscience (PacBio) (Eid, et al., 2009) and Oxford Nanopore Technologies (ONT) (Branton, et al., 2008) now allow the sequencing reads to span repeats, resulting in more continuous sequences. These advances, together with the drop of price per base, have motivated the formation of large consortia, such as the Earth Biogenome Project that aims to sequence all eukaryotes on Earth (Lewin, et al., 2018). Some projects are also aiming to produce gold standard reference genomes by using scaffolding methods like HiC and Bionano, and manual finishing to generate telomere-to-telomere assemblies with less than 1 error per 10,000 bases (Chain, et al., 2009; Koepfli, et al., 2015).

Individual research groups tend to apply whole genome sequencing to either produce *de novo* assemblies or to evaluate genome variation (single nucleotide and copy number variants, SNV-CNVs) within a species. However, the overall quality of the assemblies varies considerably due to different sample quality, genome size, fragment length and sequencing depth, which result in higher number of contigs or low consensus quality. Another reason for discontinuous assemblies is the presence of repetitive regions in the genome, so the sequencing reads need to span these repeats. In addition, there are difficulties and errors that are intrinsic to the sequencing technologies, in particular for diploid genomes. Although there are currently attempts to apply algorithms to phase diploid genomes, generating a single collapsed consensus is considered to be a more pragmatic(Garg, et al., 2018).

As the outcomes from *de novo* assemblies are usually not perfect and very heterogeneous, post-assembly correction and processing are critical steps. These steps have traditionally taken as much time and manual effort as prior sequencing and assembly, but with the advent of Illumina sequencing several steps can now be automated with tools like PAGIT (Swain, et al., 2012). For long read assemblies, improving the sequence quality (polishing) is particularly needed, as the ONT and PacBio technologies cannot capture correctly repeats of 1-3 base pairs (homopolymer tracks and di-trimer repeats)(Koren, et al., 2019; Watson and Warr, 2019). This issue is more pronounced for organisms with skewed base compositional bias, but it affects the accuracy of the predicted gene models in all organisms. Therefore, tools such as iCORN2 (Otto, et al., 2010) and Pilon (Walker, et al.,

2014) have been specifically developed to map Illumina short reads to correct genome assemblies, small errors and frameshifts. While many assembler software exist, pipelines to streamline finishing tools are still limited (Swain, et al., 2012). To our knowledge, Assemblosis (Korhonen, et al., 2019) is one of the few current pipelines that automatically assembles long reads using Canu and corrects using Pilon.

In this study, we show that there is a need to generate novel approaches to streamline the automatic finishing process. To that end, we developed the automatic Improvement of Long Read Assemblies (ILRA), an easy-to-use pipeline that combines existing tools to clean *de novo* genome assemblies using short Illumina reads. It finds contained contigs, merges contig overlaps, reorders and renames contigs based on a reference, circularizes plasmids, polishes the sequences using iCORN2 and Illumina short reads, and can detects contamination with bacteria or host sequences. We applied the ILRA pipeline to several genomes with varying sequencing depth, median read length and sequencing approaches, including four novel *Plasmodium falciparum* genomes, a *Trypanosoma brucei* assembly by PacBio (Muller, et al., 2018), and two fungi assemblies (*Leptosphaeria* spp.) by ONT (Branton, et al., 2008). We conclude that ILRA is applicable to long read assemblies, very handily and easy to use, and thus a valuable resource that could be widely implemented by non-specialists in many ongoing and future sequencing projects, especially those focusing on smaller and/or challenging genomes such as those of *Plasmodium* spp.

METHODS

Assemblies and annotation

In order to ensure that no assembler is generating assemblies without the further need of improvements, different assembler software were compared: HGAP (Chin, et al., 2013), Canu (Koren, et al., 2017), Wtdbg2 (Ruan and Li, 2020) (all using just PacBio reads), and MaSuRCA (Zimin, et al., 2013) (combining both PacBio and Illumina short reads). As input, we used four sets of reads for three different *P. falciparum* samples (PfCO01, PfKE07 and Pf2004; Table 1 and Supplementary Table 1). Some of them posed different challenges: DNA for the PfKE07 sample was amplified (WGA) and the same library for the PfCO01 sample was sequenced using two different PacBio sequencers, the RSII and the Sequel.

HGAP was run with its default parameters, setting the genome size to 23.5 Mb. We used version 3 for all assemblies, except for PfCO01 (reads from Sequel) where version 4 was used. Canu v1.8 was run with parameters -pacbio-raw, corMaxEvidenceErate=0.15, genomeSize=24m. Wtdbg2 v2.4.20190417 was run with -g 24m, -x rs for PfKE07 and the run of PfCO01 from RSII, and -x sq for the alternative run of PfCO01 from Sequel. MaSuRCA v3.3.2 was run with default parameters, and Illumina short reads (specifying mean=700 and stdev=100) were also provided. The assemblies by Wtdbg2 were preliminary polished using the Illumina short reads, following the recommendations by the authors (https://github.com/ruanjue/wtdbg2). All jobs were run on the same machine, with 40 cores.

All the assemblies in this study were annotated using the Companion (Steinbiss, et al., 2016) webserver (http://protozoacompanion.gla.ac.uk/) version 2021. We used *P. falciparum* 3D7 and *T. brucei* TREU927 as references for the corresponding assemblies. For the fungal assemblies (*Leptosphaeria* spp.), we first used the interactive Tree of Life (Letunic and Bork, 2019) to determine that, amongst the available references in Companion, *Fusarium verticillioides* is the closest fungal species, so we used it as reference for annotation. Companion default parameters were used in all cases, except for AUGUSTUS score threshold = 0.2, and taxon ID = 5833, 5691, 5022 and 220672 for *P. falciparum*, *T. brucei*, *L. maculans* and *L. biglobosa*, respectively. When available, the numbers of annotated genes and pseudogenes in the *P. falciparum* and *T. brucei* reference genomes were consulted in GeneDB (Logan-Klumpler, et al., 2012). Full statistics and information for all the assemblies in this study are in Supplementary Table 1.

Visualizations and analyses comparing the assemblies to each other and to the references were performed using Artemis and the Artemis Comparison Tool (ACT) v16.0.9 (Carver, et al., 2008). Primary assembly statistics were obtained using the software assembly-stats (https://github.com/sanger-pathogens/assembly-stats).

Comparison of iCORN2 and Pilon correction

iCORN2 v0.96 and Pilon v1.23 were evaluated for their performance and accuracy of corrections of small indels and one base pair errors in long read assemblies. As “truth set” for the experiment, the uncorrected long read assembly of *P. falciparum* (Pf3D7, available at ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/PF3K/ReferenceGenomes_Version1/Pf3D7_PacbioAssemblies/) was used by both tools and compared to the current Pf3D7 reference (version v3.1). Illumina short reads of different lengths, 75 bp to 300 bp, were evaluated. The tools were run with the following parameters: 500 bp fragment length and 5 iterations for iCORN2, and for Pilon we provided aligned reads and used the parameters --frags, --pacbio and --fix all. Two different approaches were also implemented for each tool:

1) Mapping Illumina short reads with the BWA-MEM v0.7.12-r1039 (Li, 2013) aligner with default parameters for Pilon. iCORN2 used its default aligner SMALT v0.6.2 (https://github.com/rcallahan/smalt).

2) Mapping Illumina short reads with more sensitivity using Bowtie2 v2.3.0 (Langmead and Salzberg, 2012) with parameters -X 1200 --very-sensitive -N 1 -L 31 --rdg 5,2.

To evaluate the corrections, MEGAblast v2.2.26 (Morgulis, et al., 2008) with parameters -F F -e 1e-80 -m 8 was used to analyze syntenic hits of length ≥ 10 kb, focusing on a single region in chromosome 5 (Supplementary Table 2). The results for all chromosomes were processed manually to approximately identify the same regions in the different comparisons, between the reference genome and the differentially corrected assemblies, and to evaluate the correction steps by showing final aggregated values (Table 2 and Supplementary Table 2). We used the Companion webserver version 2021 to annotate the genomes and to assess the different number of pseudogenes after the independent correction steps by both tools.

For the evaluation of the correction of a *T. brucei* genome, we used an uncorrected long read assembly (Muller, et al., 2018) and two sets of Illumina short reads that we concatenated (Supplementary Table 3). As truth set, we used the *T. brucei* Lister strain 427 2018 reference, using MEGAblast as described above. We kept the hits to the chromosomes in the reference genome larger than 10 kb and with identity > 99%. The results were processed manually to take unique alignments, prioritizing larger hits with higher bit-scores. The output for the comparison of the corrections of the *T. brucei* assembly is provided in Supplementary Table 3.

Assemblosis pipeline

Assemblosis (Korhonen, et al., 2019) was run with default parameters on two of our novel *P. falciparum* datasets (PfCO01 from Sequel and Pf2004) and compared to the results of ILRA (Supplementary Table 1).

ILRA methods

We implemented ILRA as a bash script to automatically improve long reads assemblies. Supplementary Figure 1 shows the overview of the pipeline. The full information for the assemblies in this study and the statistics for the correction by ILRA are in Supplementary Table 1. The following steps are automatically performed by ILRA:

1) **Cleaning contigs**: Contigs smaller than 5 kb are removed, as the mean read length for PacBio technologies is over 7 kb and small contigs are considered inadequate reads, for example chimeric reads. Contained contigs that fully overlap with the sequence of larger contigs are also removed, since they may represent partially phased regions of the genome or mistakes by the assembler software. The overlap between contigs is determined using MegaBLAST v2.2.26 (Morgulis, et al., 2008) with parameters -W 40 -F F -m 8 -e 1e-80. Finally, the contigs overlapping by more than 90% of their length and displaying more than 99% identity with any other contig are also removed. The names of all excluded contigs are stored in the file *Excluded.contigs.fofn*.

2) **Merging contigs**: Overlapping contigs are merged if they have an overlapping region larger than 2 kb, they display identity larger than 99% and the coverage of the Illumina short reads is 40-60% of the median coverage in the full assembly. These requisites ensure that contigs are not merged due to repeats.

3) **Ordering contigs against a reference (optional)**: The sequences are renamed and the contigs are reordered and orientated using a reference genome via ABACAS2(Assefa, et al., 2009), specifying by default minimum alignment length 1 kb, 98% identity cut-off and ABA_CHECK_OVERLAP = 0. If no reference is provided, this step is omitted and the sequences get renamed later. Here, we used the Pf3D7 reference v3.1 (Bohme, et al., 2019) for the improvement of the four *P. falciparum* assemblies. To not order contigs in polymorphic regions, we used just the core region of the genome as reference (Otto, et al., 2018). In this study, we also improved a *T. brucei* assembly by PacBio using as reference the Lister strain 427 2018 (Muller, et al., 2018), and two fungi assemblies by ONT (*Leptosphaeria maculans* Nz-T4 and *L. biglobosa* G12-14) (Dutreux, et al., 2018) using previously published genomes as references (Grandaubert, et al., 2014).

4) **Correcting homopolymer errors**: Apart from the common limitations of genome assemblies, long reads have particular shortcomings that relate to the presence of homopolymer tracks. ILRA corrects single-base discrepancies and indels using Illumina short reads via iCORN2, specifying by default 500 bp fragment length and 3 iterations. Here, we implemented a modified version of iCORN2 (v0.96). The main change we introduced is the use of Bowtie2 v2.3.0 (Langmead and Salzberg, 2012) for read mapping, in --very-sensitive mode (parameters -X 1200 --very-sensitive -N 1 -L 31 --rdg 5,2). This version has been made available in the main iCORN2 repository (http://icorn.sourceforge.net/). The tool requires Java runtime v1.7. The numbers of corrected SNPs and indels are also provided by iCORN2.

5) **Circularizing plasmids**: The genomes of organelles and extrachromosomal plasmids are circular, and the correct sequences need to be generated starting at the origin. Thus, the sequences corresponding to the mitochondria (or the apicoplast in some parasites such as *Plasmodium*, or the *T. brucei* “maxicircle” contig), are circularized by ILRA using Circlator v1.5.5 (Hunt, et al., 2015) with default parameters (command “all”). The corrected long reads provided by the assembler software and mapped to the assemblies are required by Circlator. ILRA provides the alignment using minimap2 v2.2.18 with default parameters (except for -x map-pb -H for PacBio reads and –x map-ont for ONT reads), and allows adaptation by the users, so they can specify which sequences should be circularized with inline parameters.

6) **Decontaminating contigs**: Contigs that are products of contaminations are identified by ILRA using a taxonomic classification approach by Centrifuge v1.0.4 (Kim, et al., 2016) with default parameters (NCBI nucleotide non-redundant sequences as reference and --min-hitlen 100). Decontamination is recommended prior to *de novo* assembly process, but ILRA is designed at this step to filter the contigs that are assigned to NCBI taxons representing potential contaminants. Contigs are removed using the subtraction tool Recentrifuge v1.3.1 (Marti, 2019) with default parameters, except for particular taxon IDs to be included and excluded (parameters -i and -x, respectively). The excluded contigs that are classified as contamination are recorded in the file *Excluded.contigs.fofn*. To allow adaptation by the users, the NCBI taxon IDs to keep and prioritize are provided to ILRA as an inline parameter, and the taxon IDs to be removed are provided to ILRA in the file *exclude_taxons_recentrifuge_ILRA.txt* (by default 2, 10239, 4751, 40674 and 81077, corresponding to Bacteria, Viruses, Fungi, Mammals and artificial sequences, respectively).

As centrifuge is time and memory consuming, ILRA can alternatively perform BLAST of the final assembly against multiple databases, such as common contaminants, vector sequences, bacterial insertion sequences or ribosomal RNA genes. The aim is to mask some regions, together with renaming and reformatting some contigs if needed (*e.g.,* adding the cell location), so the requirements of the most popular online databases (*i.e.* DDBJ/ENA/Genbank) are fully fulfilled. Thus, the ILRA-corrected assemblies may be directly uploaded to databases with minimal need of user input.

7) **Gathering the assembly statistics**: To assess the completeness of the chromosomes, telomere-associated sequences and telomere repeats are counted. The number of chromosomes with both telomeres attached is also estimated. The sequences to analyze within the telomeres are set by default but can be also provided to ILRA as inline parameters, to allow adaptation by the users. To evaluate the quality of the assemblies, ILRA also reports general statistics pre- and post-correction, including sequencing depth, read lengths, contig sizes, GC-content, assembly sizes, N50 values or number of gaps. Optionally, if a reference genome and an annotation file (GFF format) are provided, the software QUAST v5.0.2 (Mikheenko, et al., 2018) is also used to compute various metrics and

plots, such as structural variants compared to the reference, mis-assemblies or mismatches. Here, for assessing the quality of the ILRA-corrected assemblies, we downloaded the annotation for *P. falciparum 3D7* and *T. brucei* Lister strain 427 2018 from GeneDB and the SiegelLab.

Sequence datasets

To develop and test our pipeline, we used three species with different sets of reads: (Supplementary Table 1)

1) Through the Pf3K project reference (https://www.malariagen.net/projects/Pf3k), we generated several *Plasmodium falciparum* read sets using different library generation approaches, including WGA, Pacbio Sequel and PacBio RSII. More information on ethical and sequencing information are in the Supplementary Material.
2) *T. brucei* uncorrected PacBio long reads assembly (Muller, et al., 2018).
3) *L. maculans* Nz-T4 and *L. biglobosa* G12-14 ONT long reads assemblies (Dutreux, et al., 2018).

RESULTS

In this study, we first show that independently of the choice of assembler software, polishing is necessary. Therefore, we compared two tools to correct homopolymer tracks, Pilon and iCORN2, with the aim of choosing one to include in our pipeline to improve draft long read assemblies. Next, we implemented ILRA, compared it to Assemblosis and applied it to seven assemblies of PacBio and ONT reads from three species.

Isolates	**PfCO01**	**PfCO01**	**PfKE07**	**Pf2004**
Library preparation	Standard	Standard	Whole Genome Amplification	Standard
Sequencer machine	RSII	Sequel	RSII	RSII
Number of contigs	54	232	575	20
Size (Mb)	24.29	23.99	25.67	23.50
Number of genes	6,619	5,976	6,432	5,744
Number of pseudogenes	210	1,062	742	215

Table 1: Overview of *P. falciparum* assemblies by HGAP. Information of the *P. falciparum* datasets used in this study and statistics for the assembly process by HGAP. Standard - normal DNA preparation.

Comparison of assembler software

First, we evaluated the impact of the use of different assemblers on the number of contigs and of the presence of potential frameshifts due to homopolymer errors. We used PacBio reads from RSII and Sequel of different qualities (Supplementary Table 1), and one dataset (PfKE07) whose RSII reads were generated from a library subjected to Whole Genome Amplification (WGA). The four sets of *P. falciparum* reads were assembled with four different tools (HGAP, Canu, MaSuRCA and Wtdbg2) to determine the optimum algorithm (see Methods). The differences between the top three assemblers (HGAP, Canu and MaSuRCA) were minimal (Supplementary Table 1). We chose HGAP as the approach that generated the best results, which is also officially supported by PacBio as part of SMRT Link and was used to successfully generate assemblies from several *P. falciparum* isolates in a previous study (Otto, et al., 2018). Table 1 shows the information and genome statistics for the assemblies obtained using HGAP. Results indicate that the number of contigs of the best assemblies (up to 575) was much higher than the expected (14 chromosomes, 1 Mitochondrion and 1 Apicoplast),

and contained several frameshifts (210-1,062 pseudogenes), which highlighted the need for further improvement.

Read length	Tool	Indels	SNPs	Pseudogenes
75bp	Pilon	26,037	1,973	531
	iCORN2	9,768	1,961	164
100bp	Pilon	17,905	1,879	255
	iCORN2	6,402	2,235	143
300bp	Pilon	19,473	1,749	197
	iCORN2	4,485	3,919	141

Table 2: Comparison of Pilon with iCORN2. A Pf3D7 PacBio assembly (HGAP) was corrected with Pilon and with 5 iterations of iCORN2 (mapping of Illumina short reads by Bowtie2 for both tools). The results were compared against the *P. falciparum* 3D7 reference using MegaBLAST to obtain the number of differences between the reference and corrected sequence. The ground truth of pseudogenes is 155.

Comparison of methods for the correction of homopolymer tracks

The excessive number of annotated pseudogenes in the HGAP *P. falciparum* assemblies (up to 1,062) compared with 155 in the curated 3D7(Otto, et al., 2018) (Supplementary Table 1), points to the importance of polishing, as these pseudogenes are probably due to sequencing errors. Homopolymer tracks in long read sequencing technologies are known to give rise to frameshifts, leading to the annotation of truncated gene models and pseudogenes. This is extreme for example in the case of the *P. falciparum* genome. Due to its low GC content (around 19%) tracks of A's and T's of over 15 bp in length are frequent. Thus, frameshift errors occur because long read technologies have issues detecting the correct number of bases in homopolymer tracks or AT repeats. Figure 1 shows an example of a frameshift error in a *P. falciparum* gene model due to the presence of an homopolymer track.

We therefore compared the two mostly used packages for the task of correcting homopolymer tracks, namely iCORN2 and Pilon, by correcting a long read assembly of the *P. falciparum* 3D7 clone and comparing it to the 3D7 reference (Bohme, et al., 2019) (Table 2 and Supplementary Table 2). When correcting, we also used Illumina short reads of different lengths to determine the impact of read length. Both tools were run using alternative aligners in the mapping step and we evaluated the performance and accuracy of the correction by assessing the number of changes when comparing the corrected sequences to a reference genome (see Methods for details). We observed that the iterative approach of iCORN2, when using the Bowtie2 aligner, recognizes and corrects more errors than Pilon (Table 2 and Supplementary Table 2), but at the cost of being more than 10 times slower. After the

correction step, the annotation contained less pseudogenes (Supplementary Table 2). For instance, when correcting with 75 bp Illumina short reads, 531 pseudogenes were annotated in the Pilon-corrected assembly, whereas 164 pseudogenes were annotated in the same assembly after the iCORN2 correction, which is closer to the 155 pseudogenes from the reference. Overall, around 2-4 times more indels were also corrected with iCORN2 (Supplementary Table 2).

Figure 2 shows some examples of the differential correction of frameshifts in the *P. falciparum* 3D7 assembly by Pilon and iCORN2. As expected, the number of corrections also increased with the read length independently of the program, since longer reads can be aligned more accurately and span larger indels. Interestingly, the number of SNPs increased with iCORN2, especially when using longer Illumina short reads (300 bp). Further, 3,919 SNPs and 4,485 indel errors were still uncorrected by iCORN2 compared with the Pf3D7 reference (Table 2 and Supplementary Table 2). As we cannot determine whether this is due to errors in the correction process, the use of different DNA batches or the presence of errors within the reference, we just used the number of pseudogenes remaining after correction as the best quality metric. Finally, we confirmed that iCORN2 is superior to Pilon by repeating this approach on a *T. brucei* uncorrected assembly. When compared with the *T. brucei* Lister strain 427 2018 reference, around 20% less indels were found in the iCORN2-corrected sequences than in the Pilon-corrected ones and less pseudogenes were annotated (Supplementary Table 3). Based on this better performance, we chose to incorporate the iCORN2 software to the ILRA pipeline.

Automatically correction of *de novo* genome assemblies by ILRA.

As described in more detail in the Methods (Supplementary Figure 1), our pipeline first cleans short contigs and contained contigs. Overlaps are then found with a novel method based on merging only if the Illumina coverage is around half of the expected read depth. If a reference sequence exists, the new contigs are also orientated against the reference and renamed. Next, homopolymer errors are corrected using the iCORN2 software as described above, and plasmids are circularized. Finally, the assembly and contigs get decontaminated (for example host or bacterial contamination) and basic statistics are generated, such as the number of potentially complete chromosomes. ILRA is programmed in Bash and uses programs that are mostly coded in Perl.

First, we tested the ILRA pipeline on a *T. brucei* uncorrected long read assembly (Muller, et al., 2018). The statistics are included in Table 3 and Supplementary Table 1. As expected, the ILRA pipeline successfully corrected *T. brucei* sequences, improving contiguity from 1,232 to 569 contigs (reference genome contiguity = 317) and genome size from 65.5 Mb to 57.2 Mb (reference genome size = 50.1 Mb). The number of annotated pseudogenes also decreased by 1,358 (Table 3/ Supplementary Table 1).

Isolates	*T. brucei*	*L. maculans Nz-T4*	*L. biglobosa G12-14*	PfCO01 R	PfCO01 S	PfKE07 R	Pf2004 R
Pre-ILRA							
#Contigs	1,232	288	156	54	232	575	20
#Pseudogenes	8,074	540	498	210	1,062	742	215
ILRA corrections							
#Excluded contigs <5 kb	94	0	0	4	15	72	0
#Excluded contained contigs	254	0	0	16	18	81	0
#Excluded contigs not covered and merged	22	0	0	7	26	23	2
#Contigs merging and reference ordering	240	186	49	5	108	202	0
#Excluded contigs contamination	53	32	0	3	1	0	0
#Contigs (final)	569	70	107	19	64	197	18
#Pseudogenes (final)	6,716	422	415	95	246	311	115

Table 3: Overview of ILRA improvements step by step of the seven datasets used in this study. Pre-ILRA are the statistics from HGAP and it is shown how each step of the ILRA pipeline improves the assemblies (R – RSII, S – Sequel).

Next, we demonstrated that ILRA is also applicable to *de novo* genome assemblies from ONT reads. Here, we ran ILRA on two recently published fungal genomes (Dutreux, et al., 2018) and observed that ILRA produced significant improvements in contiguity, annotation of pseudogenes, and corrections of thousands of SNPs and indels. While no good reference genomes are available for *Leptosphaeria* spp, contiguity for a *L. maculans* strain Nz-T4 assembly was improved from 288 to 70 contigs and the annotated pseudogenes decreased from 540 to 422, which is closer to the number

reported in previous assemblies (Table 3, Supplementary Table 1). Similar improvements were also observed for a *L. biglobosa* G12-14 strain assembly, with 31% less contigs and 17% less annotated pseudogenes (Table 3, Supplementary Table 1).

In this study, we also generated novel *P. falciparum* assemblies from three *P. falciparum* isolates with diverse origin: Colombia (PfCO01), Kenya (PfKE07) and Ghana (Pf2004). Table 1 and Table 3 show that the quality of the assemblies obtained using HGAP varies significantly between isolates. For example, the amount of contigs ranged from 20 to 575 and the annotated pseudogenes from 210 to 1,062. These differences can be attributed to contamination, different median read lengths, different sequencing technologies for PfCO01, or the library preparation protocol, which in the case of PfKE07 was based on whole genome amplification (WGA). We observed that the genomes coming from PacBio RSII reads (Table 1), were finally assembled into a range of 18-197 contigs (median=19). These assemblies clearly benefited from the application of ILRA automatic correction, with lower contiguity (previously, range of 20-575 contigs, median=54) and lower number of annotated pseudogenes (Table 1). For example, ILRA allowed for the improved contiguity and the correction of 100 genes incorrectly predicted to be pseudogenes in the Pf2004 sample, which finally assembled into just 18 contigs and had 5,761 genes and 115 pseudogenes annotated (20 contigs, 5,744 genes and 215 pseudogenes before ILRA, Table 1). Overall, more than 5,600 genes were annotated in all cases, and contiguity and number of annotated pseudogenes were always improved after ILRA (Table 1, Supplementary Table 1). Moreover, for Pf2004 several contigs contained terminal telomeric repeats, which could indicate fully assembled chromosomes, as 11 out of the 18 contigs of the ILRA-corrected Pf2004 assembly had both telomeres attached (Table 1, Supplementary Table 1). Despite the general improvements of the ILRA pipeline, there were also some assemblies of lower quality, such as PfKE07. For this assembly, the number of gaps was still high (Supplementary Table 1) and we observed artefacts and mis-assemblies due to the WGA preparation. This may be due to the polymerase switching between strands during the amplification step, which results in inverted chimeras generating miss-assemblies that our pipeline was not able to address. Figure 3 displays a case example of an error due to WGA and chimeric reads.

For the last performance test, we sequenced the library from the *P. falciparum* Colombian sample (PfCO01) both with the PacBio RSII and the PacBio Sequel chemistries. Of note, the Sequel was one of the first runs ever performed and since then the chemistry has been continuously improved. The assembly with the PacBio RSII reads was of considerably better quality, with approximately 25% of the Sequel contigs before ILRA. After correction by ILRA, the sequences coming from Sequel reads assembled into 64 contigs, while the assembly from RSII reads was composed of 19 contigs (Table 1, Supplementary Table 1). To note, the same library was sequenced with both machines, but the mean read length was longer in the RSII (9,413 versus 7,668), but the read depth was higher in the Sequel run (198 versus 168, Supplementary Table 1). Another

improvement by ILRA in this assembly was the identification and removal of some *Mycoplasma* contamination that also caused an excessive number of annotated genes and pseudogenes in both PfC01 assemblies (Table 3).

Finally, for two *P. falciparum* novel genomes we also compared ILRA to Assemblosis (Korhonen, et al., 2019), an alternative software for the assembly and correction of sequences. Assemblosis performed well, but the assemblies are 1.4 and 0.6MB shorter. Never the less, ILRA generated a more continuous assembly and produced less pseudogenes for both PfCO01 from RSII reads and Pf2004 (19 versus 30 contigs and 95 vs 160 pseudogenes for PfCO01, and same contiguity, 18 contigs, but 115 vs 170 pseudogenes for Pf2004, Supplementary Table 1).

DISCUSSION

With the advent of long reads technologies, the rapid drop of costs and the development of associated algorithms for the analysis, genome sequencing has become more popular and accessible for many laboratories worldwide. However, we show that as expected, polishing of the assemblies is essential to improve the quality of the final genomes and is likely to pose a challenge to laboratories without deep bioinformatics knowledge. Polishing includes steps such as the removal of small contigs, circularization of mitochondria, naming of sequences and addressing errors due to homopolymer tracks and the resultant frameshifts. Here, we develop ILRA, which is a pipeline that combines existing and new tools performing these post-sequencing steps in a completely integrated way, providing fully corrected and ready-to-use genome sequences. In this study, we compare its application and performance on existing assemblies from different organisms, as well as novel genome assemblies obtained using various sequencing methods. We show that in all cases ILRA significantly improved the outcome, even if the quality of the original assembly was low (Table 3). Interestingly, not many pipelines to automatically finish genomes exist. In our test we can show that ILRA is better than the test tool.

The correction of homopolymer tracks with Illumina reads is crucial when correcting long read assemblies from PacBio and ONT reads. Amongst the existing software applied for this purpose, we could clearly show that in *P. falciparum* and *T. brucei*, iCORN2 supersedes other tools (Table 2), thanks to its iterative nature. Even if the runtime is high, it seems an iterative approach should always be implemented when improving long reads genome assemblies, in order to make use of the full capabilities of short reads. Thus, the sequences can be iteratively corrected until enough errors are fixed and Illumina short reads are mapping better. In fact, despite Pilon not supporting this feature natively, an iterative approach with multiple rounds of correction with this software has recently been used with some success by others (Koren, et al., 2017; Naquin, et al., 2018; Tan, et al., 2018). Here, we incorporated an updated version of iCORN2 within the ILRA pipeline. We also tested this

correction step using Illumina short reads of different lengths (Supplementary Table 2) and consistently observed a better performance of iCORN2, with best corrections using 75 bp and 100 bp short reads. Although with longer reads (300 bp) we showed an improvement of corrected indels, we also observed an increase of single base errors (SNPs). Thus, this increase could be an artefact due to the higher error rate of reads after homopolymer tracks, or genuine differences between the reads and the current genome reference. The fact that there is not really a ground truth for a perfect genome highlights the outstanding dilemma that for some reference genomes it is difficult to determine the final consensus sequence.

To further test ILRA, we applied it to several novel and published genomes, including parasites and fungi, both assembled from PacBio and ONT technologies. Overall, we show that IRLA always improves the contiguity of the assemblies and genome sizes, as well as reducing the number of wrongly assigned pseudogenes. Therefore, ILRA represents an excellent tool for the community to automatically improve long read genome assemblies. However, due to intrinsic limitations in the homopolymer correction step performed by iCORN2 and in the reference-based reordering and renaming by ABACAS2, we would recommend the use of ILRA for the improvement of genomes up to the size of 300 Mb.

We have also explicitly shown that, independent of the choice of assembler software, post-assembly polishing steps are always needed. Our results show that HGAP and Canu are the best assemblers for *Plasmodium* sequences, and that for *P. falciparum de novo* assemblies our approach based on HGAP and ILRA supersedes. Apart from benefiting from the polishing by ILRA, our *P. falciparum* sequences are also case examples illustrating the importance of the initial selection of the sequencing technology. For instance, we observed differences in the PfCO01 assembly from reads by PacBio Sequel or PacBio RSII technologies, or despite achieving great improvement and better quality post-ILRA, the PfKE07 assembly was always too fragmented. The reason is that the WGA step includes artefacts (Figure 3). WGA may be required to increase the amount of DNA for sequencing, but PacBio recently generated new methods to obtain long reads from 100 ng of DNA (Kingan, et al., 2019), which could render WGA methods obsolete if both high molecular weight DNA and the required expertise is available. However, this is not always possible due to the frequent difficulties to obtain adequate samples when DNA is limited or contaminated by host material. The four *de novo P. falciparum* assemblies from field isolates, generated by HGAP and automatically corrected by ILRA in this study, are now publicly available. Together with the ILRA pipeline, they are likely to be valuable references and resources for future studies.

CONCLUSIONS

Although long reads technologies now make possible to generate nearly perfect *de novo* genome assemblies from in principle any organism DNA, consensus sequences will still need

polishing and correction of homopolymer errors. Further, in many cases it is not possible to assemble reads at the chromosome level due to limited amount of DNA, low DNA quality or host contamination. In all these cases, the ILRA pipeline is an easy-to-use and accessible tool for any laboratory without deep bioinformatics knowledge, which automatically performs several polishing steps and successfully improves assemblies, making them more continuous and decreasing the number of wrongly assigned pseudogenes.

Software and Data availability

ILRA can be downloaded from GitHub: https://github.com/ThomasDOtto/ILRA. The code is also available as preinstalled virtual machine. Accession numbers of the published data are, *T. brucei*: ERR1795268/ SRR5466319, *L. maculans* NzT4: PRJEB24469, *L. maculans* G12-14: PRJEB24467, and *P. falciparum* reads: ERS037841, ERS001369 and ERS557779, for the 75bp, 100bp and 300 bp, respectively. Novel data can also be found on online databases (accession numbers for long reads, short reads and assemblies, respectively), Colombian (PfCO01, RSII long reads): ERS2460039, ERS1746432, SAMN18287201, Colombian (PfCO01, Sequel long reads): ERS2460039, ERS1746432, SAMN18287200, Kenyan (PfKE07) ERS2026796, ERS166385, SAMN18287199, and Ghanaian (Pf2004) ERS1412916, ERS1306150, SAMN18287202. The assemblies can also be found on the IRLA GitHub page.

ACKNOWLEDGMENTS

We would like to thank the patients who contributed samples and the health workers who assisted with the sample collections. We would also like to thank staff from the Illumina Bespoke Sequencing Team at the Wellcome Sanger Institute for their contribution. This paper is published with permission from the Director of Kenya Medical Research Institute (KEMRI).

FUNDING

This work was supported by the Wellcome Trust [098051, 104111/Z/14/ZR], E.G.-D. is funded by the Spanish Ministry of Science and Innovation grant no. PID2019-111109RB-I00. J.L.R is funded by a Severo Ochoa Fellowship (BES-2016-076276). D.F.E and J.D.E-P are founded by Colciencias, call 656-2014 "EsTiempo de Volver" award FP44842-503-2014 and “programa jovenes investigadores” special cooperation 552-2015, respectively.

CONFLICT OF INTEREST

None declared.

Figure legends

FIGURE LEGENDS

Figure 1: Frameshifts and homopolymer tracks in a long read genome assembly.

Artemis visualization of a PacBio genome assembly (bottom panel) and the aligned Illumina short reads (top panel, horizontal blue bars), with reads mapping to the forward strand on top, and to the reverse below. Sequencing errors in the Illumina short reads are marked with vertical light red lines. A homopolymer track of 17 A's is highlighted in yellow. The quality of the reads drops after the homopolymer, and accordingly it can be seen that reads on the forward strand have just few sequencing errors, but after the homopolymer track the error rate is high. As a homopolymer track is not sequenced correctly, it generates a frameshift and therefore makes a gene model to be wrongly annotated as a pseudogene. In the bottom panel, the two light blue boxes represent exons that due to the indel are split into two. *Ad initio* gene finders could try to build an intron here (loosing exon sequence) or to generate a pseudogene. In the zoom-in visualization (right), the dark red vertical lines in the aligned Illumina short reads point to bases that are missing from the short repetition in the assembly, resulting in the homopolymer track causing the frameshift.

Figure 2: Differential frameshifts correction by Pilon and iCORN2.

ACT visualization of a section of the Pf3D7 reference genome, the corresponding section of an uncorrected *P. falciparum* 3D7 PacBio genome assembly, and the Pilon-corrected and iCORN2-corrected sequences. Syntenic regions (BLAST) are indicated in gray bars between the reference and the uncorrected assembly. Annotated genes in the reference are colored. Red squares mark the frameshifts within Open Reading Frames (ORFs) in the uncorrected genome sequences. These are differentially processed by Pilon and iCORN2 and iCORN2 corrects more frameshifts than Pilon. Green squares mark the correct and successively corrected ORFs, which based on the reference could produce proper gene models instead of an excessive and incorrect annotation of pseudogenes.

Figure 3: Whole Genome Amplification errors in the PfKE07 assembly.

A. Schematic error of WGA. DNA gets amplified (i), but then the polymerase strand switches and generates the reverse strand (ii). This generates a chimeric read that generates mis-assemblies.

B. These chimeric reads generate assembly errors, as seen in an ACT view. The top part of the reference genome (gray arrow) is duplicated in the WGA amplified genome. The assembly errors generally occur as contig end, so gaps are generated. Syntenic regions (BLAST similarity hits) when comparing to the reference genome are indicated in gray. Miss-assemblies (Inverted similarity hits) are indicated in black.

Supplementary Figure 1: IPA pipeline.

Descriptive flowchart summarizing the steps implemented in the IPA pipeline: contigs filtering, reordering and renaming, iterative error correction, circularization, decontamination and evaluation of the genome assemblies.

REFERENCES

Assefa, S., *et al.* ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 2009;25(15):1968-1969.

Bohme, U., *et al.* Progression of the canonical reference malaria parasite genome from 2002-2019. *Wellcome Open Res* 2019;4:58.

Branton, D., *et al.* The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008;26(10):1146-1153.

Carver, T., *et al.* Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. *Bioinformatics* 2008;24(23):2672-2676.

Chain, P.S., *et al.* Genome Project Standards in a New Era of Sequencing. *Science 326:5950* 2009:236-237.

Chin, C.S., *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;10(6):563-569.

Dutreux, F., *et al.* De novo assembly and annotation of three Leptosphaeria genomes using Oxford Nanopore MinION sequencing. *Sci Data* 2018;5:180235.

Eid, J., *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009;323(5910):133-138.

Garg, S., *et al.* A graph-based approach to diploid genome assembly. *Bioinformatics* 2018;34(13):i105-i114.

Grandaubert, J., *et al.* Transposable element-assisted evolution and adaptation to host plant within the Leptosphaeria maculans-Leptosphaeria biglobosa species complex of fungal pathogens. *BMC genomics* 2014;15:891.

Hunt, M., *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 2015;16:294.

Kim, D., *et al.* Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;26(12):1721-1729.

Kingan, S.B., *et al.* A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing. *Genes (Basel)* 2019;10(1).

Koepfli, K.P., *et al.* The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* 2015;3:57-111.

Koren, S., *et al.* Reply to 'Errors in long-read assemblies can critically affect protein prediction'. *Nat Biotechnol* 2019;37(2):127-128.

Koren, S., *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;27(5):722-736.

Korhonen, P.K., *et al.* Common workflow language (CWL)-based software pipeline for de novo genome assembly from long- and short-read data. *Gigascience* 2019;8(4).

Langmead, B. and Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357-359.

Letunic, I. and Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47(W1):W256-W259.

Lewin, H.A., *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America* 2018;115(17):4325-4333.

Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013;1303.3997v1

Logan-Klumpler, F.J., *et al.* GeneDB--an annotation database for pathogens. *Nucleic Acids Res* 2012;40(Database issue):D98-108.

Marti, J.M. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Comput Biol* 2019;15(4):e1006967.

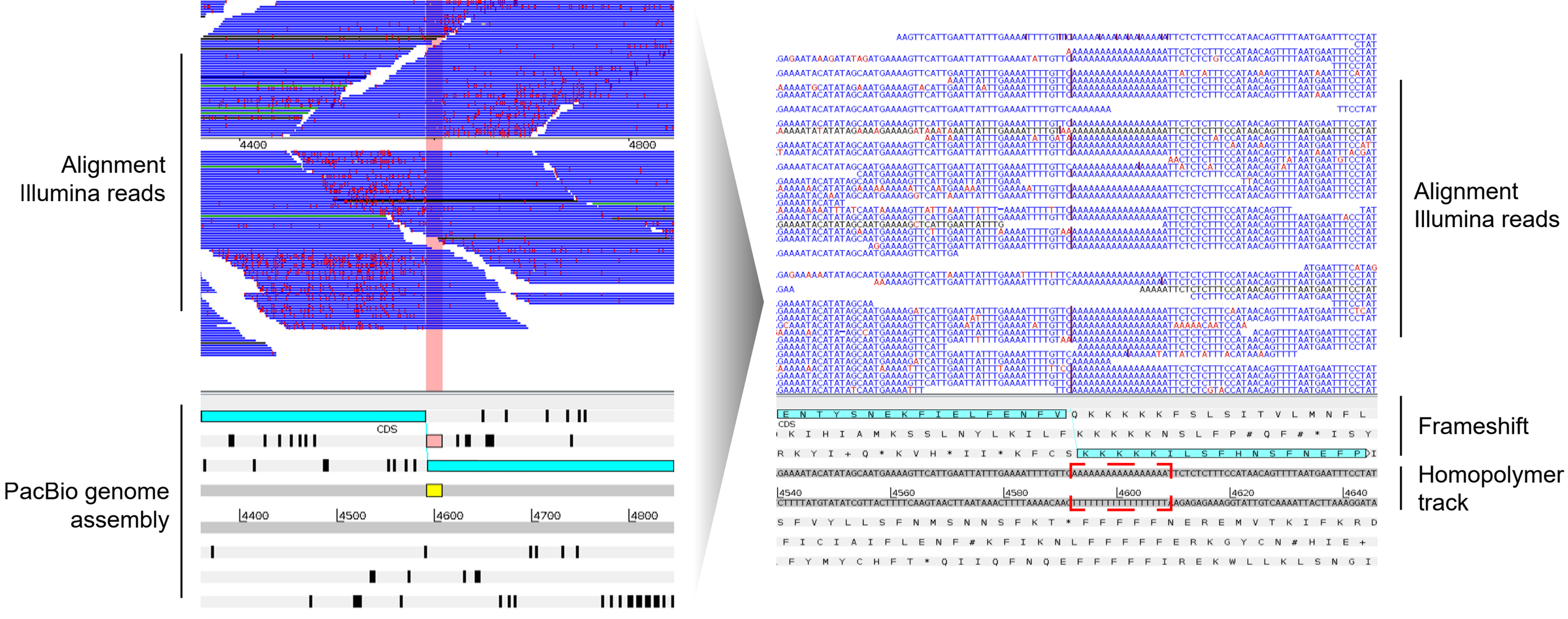
Marx, V. Long road to long-read assembly. *Nat Methods* 2021;18(2):125-129.

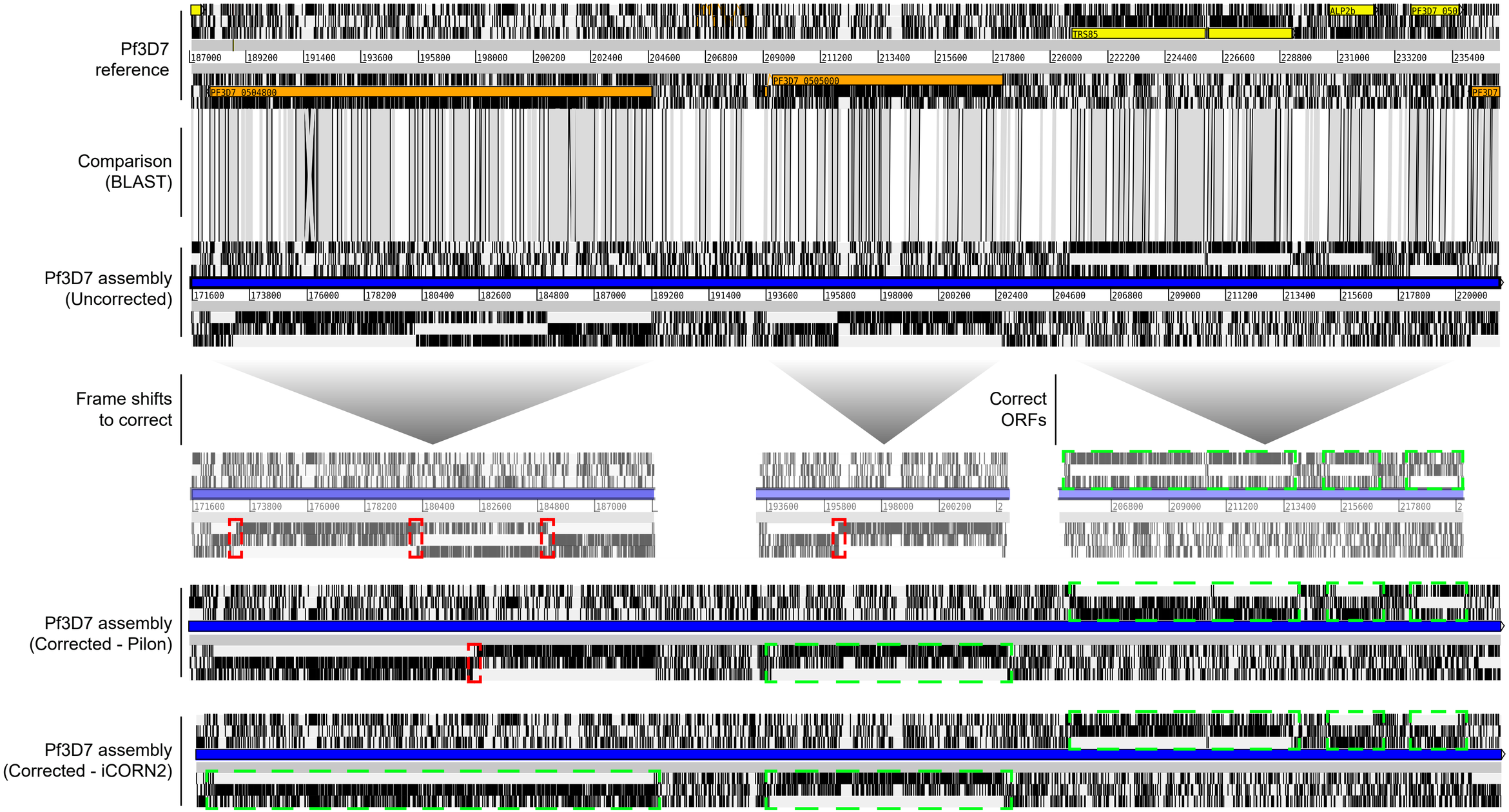
Mikheenko, A., *et al.* Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018;34(13):i142-i150.

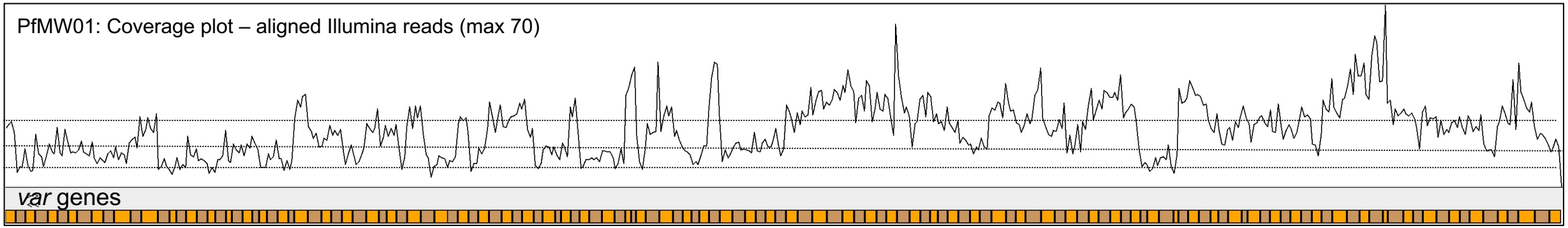
Morgulis, A., *et al.* Database indexing for production MegaBLAST searches. *Bioinformatics* 2008;24(16):1757-1764.

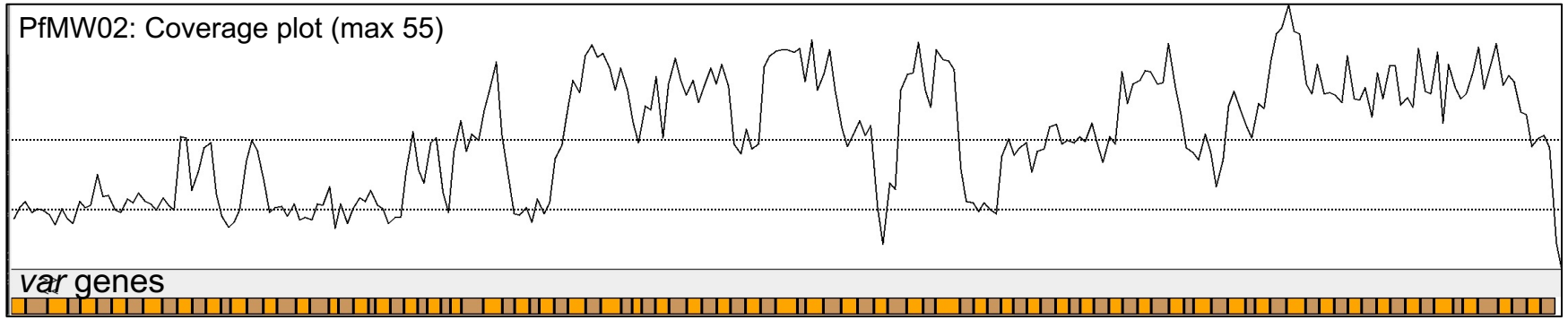
Muller, L.S.M., *et al.* Genome organization and DNA accessibility control antigenic variation in trypanosomes. *Nature* 2018;563(7729):121-125.

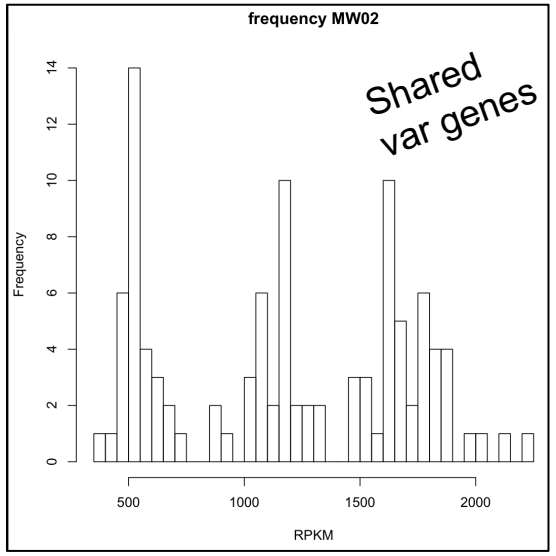
Naquin, D., *et al.* Complete Sequence of the Intronless Mitochondrial Genome of the Saccharomyces cerevisiae Strain CW252. *Genome Announc* 2018;6(17).
Otto, T.D., *et al.* Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. *Wellcome Open Res* 2018;3:52.
Otto, T.D., *et al.* Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* 2010;26(14):1704-1707.
Ruan, J. and Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;17(2):155-158.
Steinbiss, S., *et al.* Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* 2016;44(W1):W29-34.
Swain, M.T., *et al.* A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nature protocols* 2012;7(7):1260-1284.
Tan, M.H., *et al.* Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. *Gigascience* 2018;7(3):1-6.
Walker, B.J., *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9(11):e112963.
Watson, M. and Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* 2019;37(2):124-126.
Zimin, A.V., *et al.* The MaSuRCA genome assembler. *Bioinformatics* 2013;29(21):2669-2677.

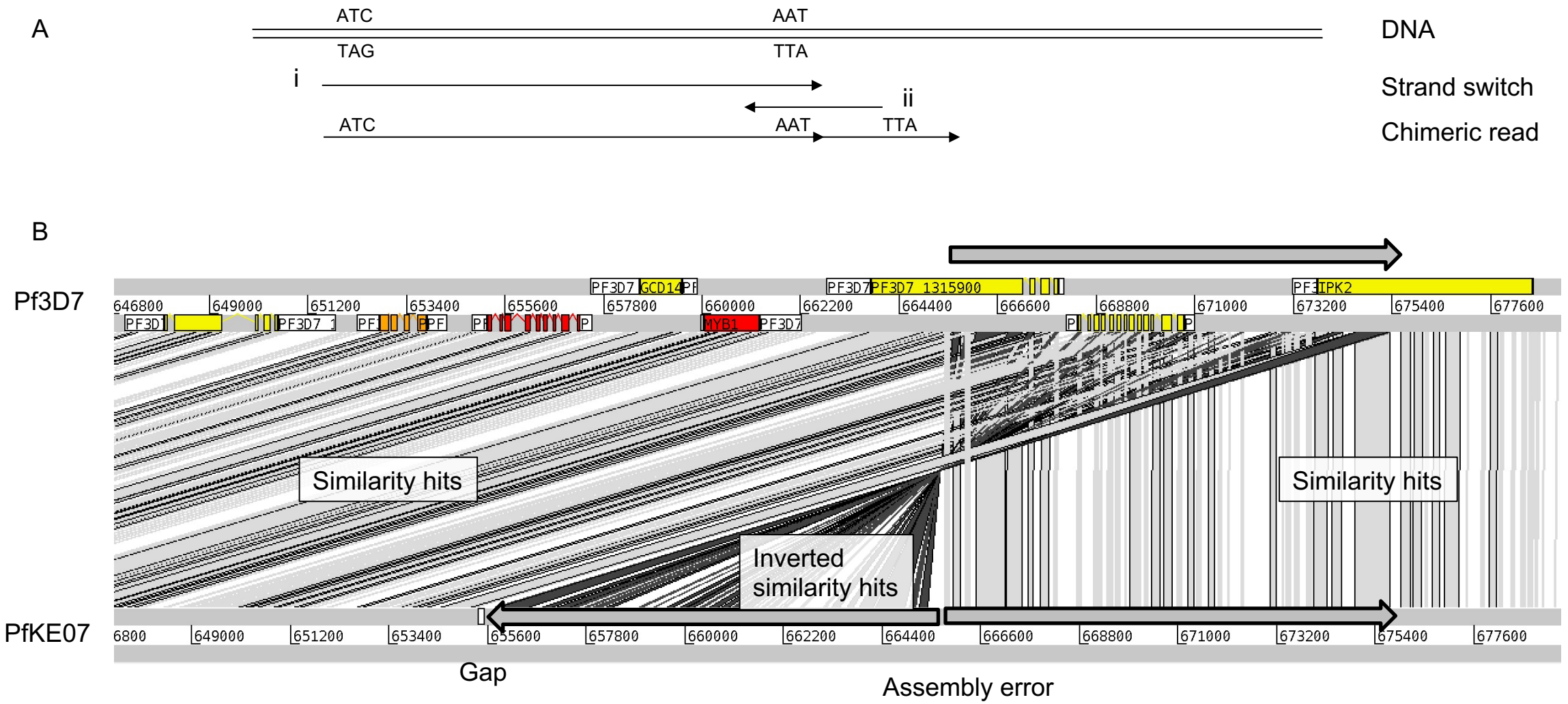

Alignment
Illumina reads
4400
4800
PacBio genome
assembly
CDS
4400
4500
4600
4700
4800
Alignment
Illumina reads
Frameshift
Homopolymer
track
4540
4560
4580
4600
4620
4640


Pf3D7 reference
Comparison (BLAST)
Pf3D7 assembly (Uncorrected)
Frame shifts to correct
Correct ORFs
Pf3D7 assembly (Corrected - Pilon)
Pf3D7 assembly (Corrected - iCORN2)
PF3D7_0504800
PF3D7_0505000
TRS85
ALP2b

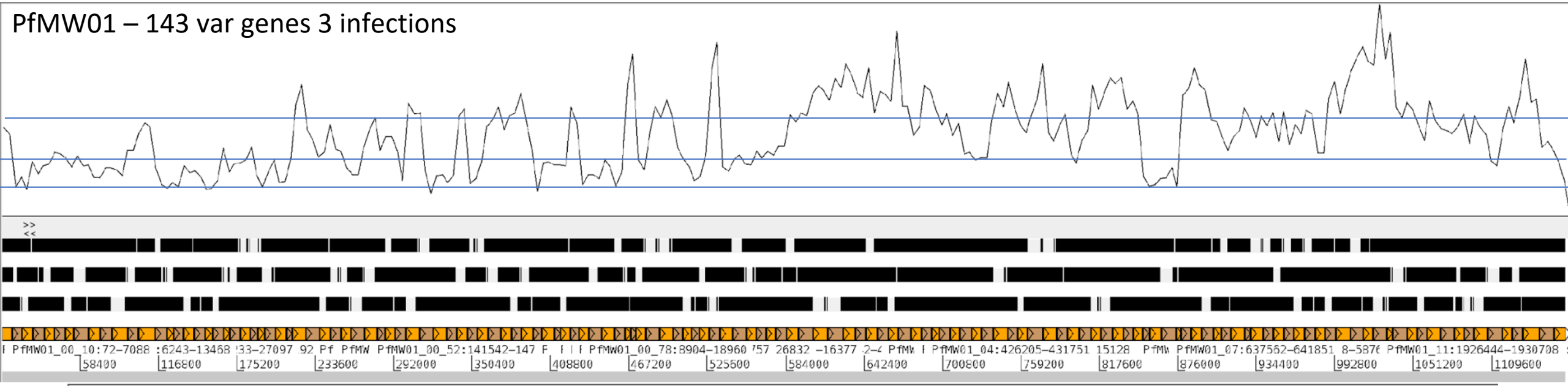
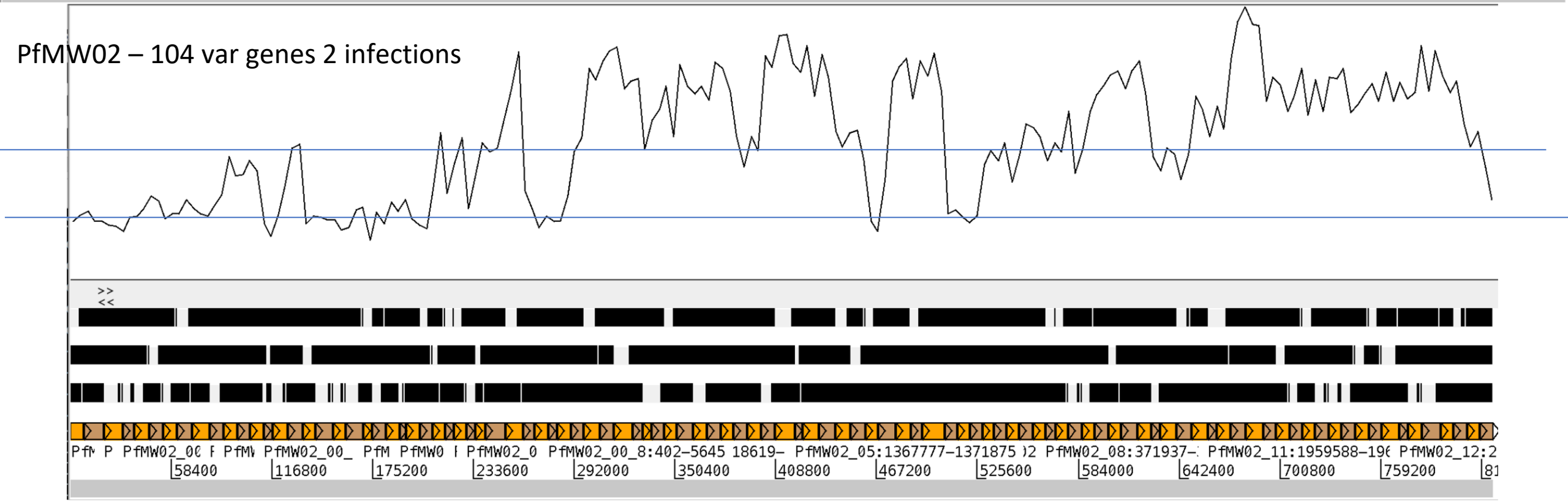

PfMW01: Coverage plot – aligned Illumina reads (max 70)
var genes


PfMW02: Coverage plot (max 55)
var genes


frequency MW02
Shared var genes
Frequency
0
2
4
6
8
10
12
14
500
1000
1500
2000
RPKM


A
ATC
AAT
DNA
TAG
TTA
i
ii
Strand switch
ATC
AAT
TTA
Chimeric read
B
Pf3D7
PfKE07
Similarity hits
Similarity hits
Inverted similarity hits
Gap
Assembly error

PfMW01 – 143 var genes 3 infections



PfMW02 – 104 var genes 2 infections