#345 (Monday, June 20, 2016)

Complete telomere-to-telomere *de novo* assembly of the *Plasmodium falciparum* genome using long-read sequencing

Shruthi Sridhar Vembar[1,2,3], Matthew G. Seetin[4], Christine Lambert[4], Maria Nattestad[5], Michael C. Schatz[5,6], Primo Baybayan[4], Artur Scherf[1,2,3], Melissa Laird Smith[4]
[1]Unité Biologie des Interactions Hôte-Parasite, Département de Parasites et Insectes Vecteurs, Institut Pasteur, Paris 75015, France; [2]CNRS, ERL 9195, Paris 75015, France; [3]INSERM, Unit U1201, Paris 75015, France; [4]PacBio, Menlo Park, CA, U.S.A.;
[5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY U.S.A.; [6]Johns Hopkins University, Baltimore, MD, U.S.A

Abstract

Sequence-based estimation of genetic diversity of *Plasmodium falciparum*, the most lethal malarial parasite, has proved challenging due to a lack of a complete genomic assembly. The skewed AT-richness (~80.6% (A+T)) of its genome and the lack of technology to assemble highly polymorphic sub-telomeric regions that contain clonally variant, multigene virulence families (i.e. *var* and *rifin*) have confounded attempts using short-read NGS technologies.

Using Single Molecule, Real-Time (SMRT) Sequencing, we successfully compiled all 14 nuclear chromosomes of the *P. falciparum* genome from telomere-to-telomere. A Hierarchical Genome Assembly Process (HGAP) was used to *de novo* assemble the *P. falciparum* genome.

This assembly accurately resolved centromeres (~90-99% (A+T)) and sub-telomeric regions, and identified large insertions and duplications in this genome that added extra genes in the *var* and *rifin* virulence families, along with smaller structural variants such as homopolymer tract expansions. Identifying the polymorphic and repetitive sub-telomeric sequences of parasite populations from endemic areas might inform the link between structural variation and phenotypes such as virulence, drug resistance and disease transmission.

Sample Preparation And Sequencing

A

SMRT Cell	Purification strategy	On-plate conc. (pM)	Total Mb	Total Number of Reads	Mean Read Length	Read Length N50
1	AMPure	200	492	47,693	12,427	15,718
2	AMPure	200	669	63,846	12,865	16,337
3	AMPure	200	632	60,912	13,122	16,375
4	Aurora	200	495	46,069	12,253	16,041
5	Aurora	200	336	33,096	12,416	16,077
6	Aurora	200	552	52,603	12,773	16,226
7	P-C*	100	750	81,949	10,823	13,943
8	P-C*	100	678	71,458	11,265	14,080
9	P-C*	100	652	68,370	11,223	14,348
Mean			***584***	***58,444***	***12,130***	***15,461***
SD*			***127***	***15,004***	***822***	***1026***

* P-C: Phenol-chloroform
** SD: Standard Deviation

B

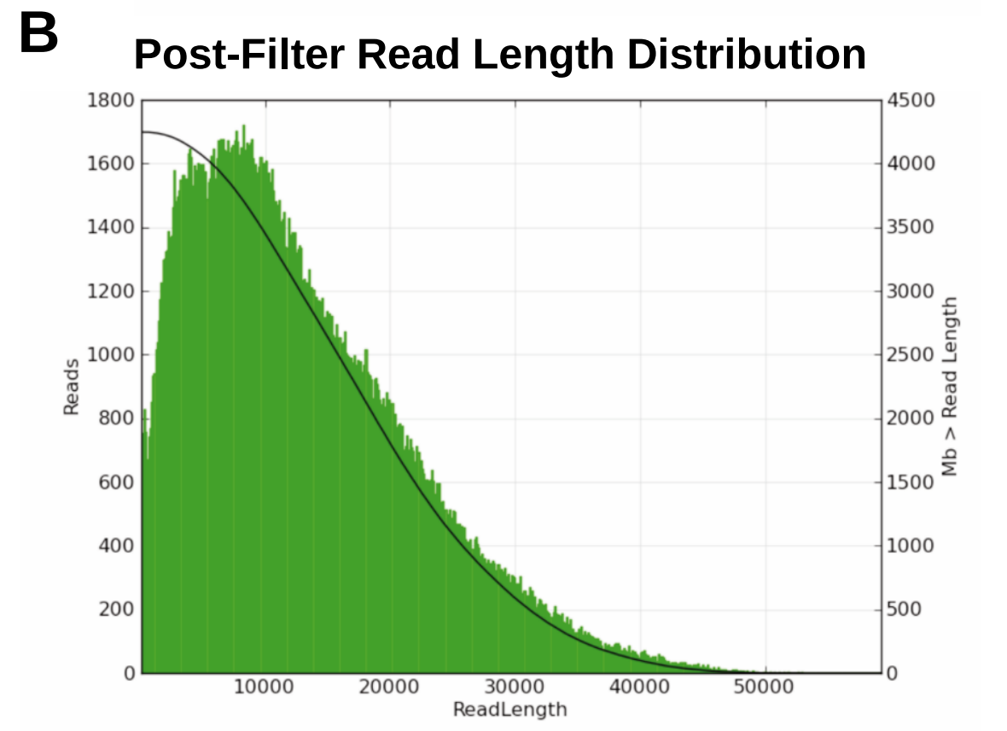


Table 1 / Figure 1. ***P. falciparum*** **SMRTbell library preparation using three genomic DNA purification methods**

High-molecular weight *P. falciparum* genomic DNA was purified in triplicate using three different methods (i) AMPure PB magnetic bead-based clean up, (ii) electrophoretic DNA extraction using the Aurora System, or phenol-chloroform extraction. Following a standard 20 kb SMRTbell library construction protocol with a 15 kb cut-off size selection, samples were sequenced on the PacBio RS II using 4 hour movies. Library size, quantity, and quality before and after size selection (data not shown) and **(A)** individual SMRT Cell analysis results were similar, so further analysis utilized pooled sequence data from all 9 SMRT Cells. **(B)** Post-filtering read length distribution. The X-axis is read length, while the Y-axes is number of reads (green columns) and megabases (Mb) greater than read length (curve). HGAP was set to use the longest 30-fold coverage of subreads for preassembly, in this case subreads longer than 23,278 bases.

REFERENCES

1. Vembar, S.S., et al., (2016) Complete telomere-to-telomere *de novo* assembly of the *Plasmodium falciparum* genome through long-read (>11 kb), Single Molecule, Real-Time sequencing. *DNA Research*. In press.
2. Volkman, SK et al., (2007) A genome-wide map of diversity in *Plasmodium falciparum*. *Nature Genetics.* 39(1), 113-119.
3. Kozarewa, I. et al., (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nature Methods.* 6(4), 291-295.
4. Sanarakoon, U. et al., (2011) High-throughput 454 resequencing for allele discovery and recombination mapping in *Plasmodium falciparum. BMC Genomics.* 12,116.
5. Krumsiek J, et al., (2007) Gepard: A rapid and sensitive tool for creating dotplots on genome scale, *Bioinformatics.* 23(8), 1026.
6. Gardner, MJ. et al., (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature.* 419(6906), 498-511.
7. Nattestad, M, Schatz, MC (2016) Assemblytics: a web analytics tool for the detection of assembly-based variants. *Bioinformatics.* In press. doi: http://dx.doi.org/10.1101/044925

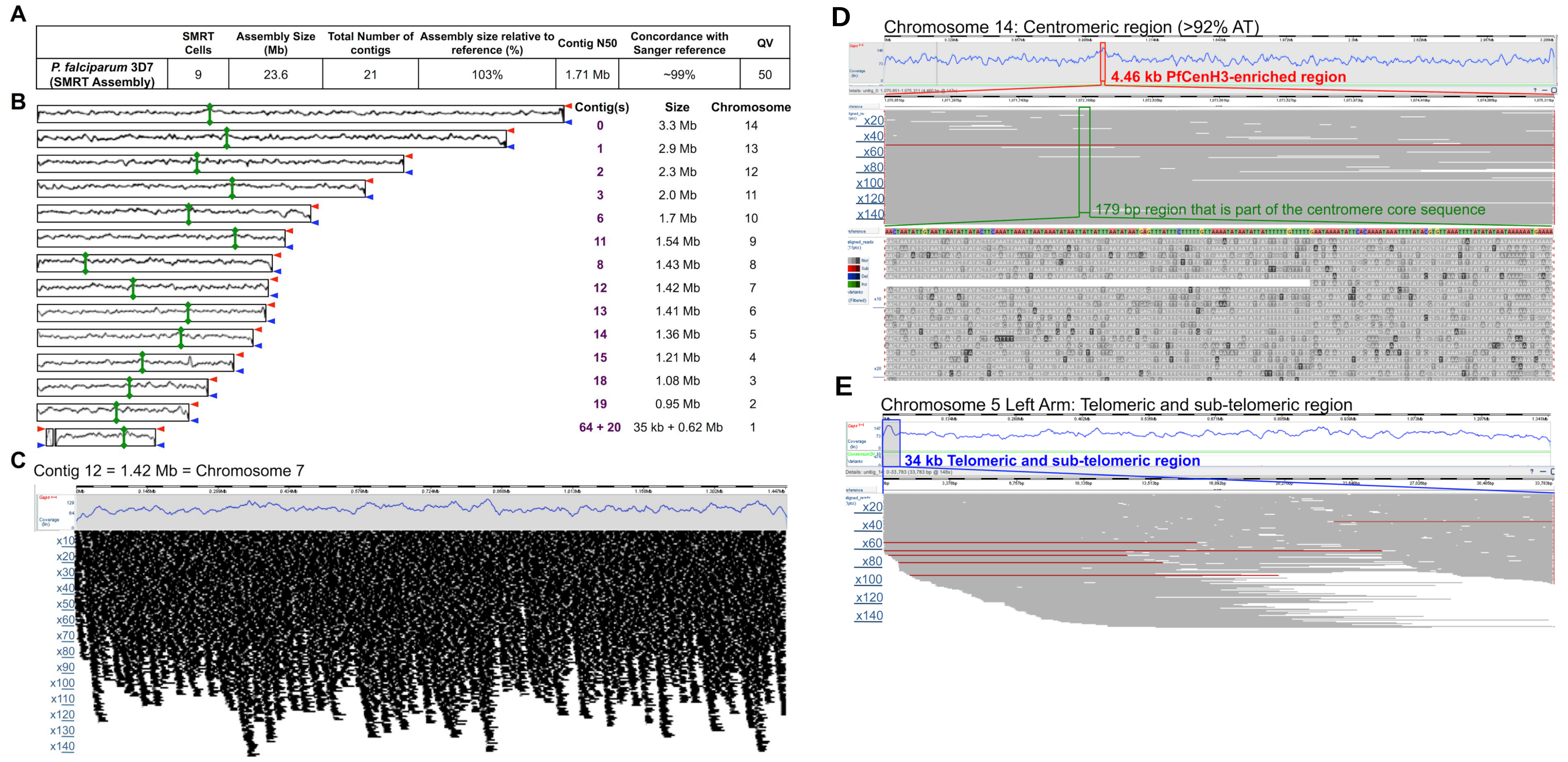
Assembly Comparison

	SANGER SEQUENCING[2]		ILLUMINA SEQUENCING[3]		454 SEQUENCING[4]		SMRT SEQUENCING (THIS STUDY)
Parasite strain	Dd2	HB3	NP-3D7-S	NP-3D7	7C126	SC05	**3D7**
Read length (bases)	600-700		36	76	3,000 (Paired end)		**12,130**
# of contigs	4,511	2,971	26,920	22,839	9,452	9,597	**21**
N50 Contig Size (kb)	11.6	20.6	1.5	1.6	3.3	3.3	**1,710**
Largest Contig (kb)	79.2	111.9	29.1	24.0	36.7	34.4	**3,290**
# of assembled bases (Mb)	19.5	23.4	19.0	21.1	20.8	21.1	**23.6**
Average Coverage	7.8-fold	7.1-fold	43-fold	64-fold	33-fold	36-fold	**94-fold**

Table 2. **Comparison of previous *P. falciparum* 3D7 *de novo* assemblies using other sequencing technologies**

Due to short read lengths and difficulties with high AT bias, other PCR-based sequencing technologies struggle to assemble contigs even as long as the mean raw reads used in this work. In contrast, this SMRT Assembly results in a nearly finished genome with a contig N50 that is itself a full chromosome and two to three orders of magnitude larger than what was possible with other sequencing approaches.

Assembly Produced Full Chromosomes



	SMRT Cells	Assembly Size (Mb)	Total Number of contigs	Assembly size relative to reference (%)	Contig N50	Concordance with Sanger reference	QV
P. falciparum 3D7 (SMRT Assembly)	9	23.6	21	103%	1.71 Mb	~99%	50

Contig(s)	Size	Chromosome
0	3.3 Mb	14
1	2.9 Mb	13
2	2.3 Mb	12
3	2.0 Mb	11
6	1.7 Mb	10
11	1.54 Mb	9
8	1.43 Mb	8
12	1.42 Mb	7
13	1.41 Mb	6
14	1.36 Mb	5
15	1.21 Mb	4
18	1.08 Mb	3
19	0.95 Mb	2
64 + 20	35 kb + 0.62 Mb	1

Figure 3. **HGAP *de novo* assembly resolves all 14 *P. falciparum* chromosomes**

(A) *P. falciparum* genome HGAP assembly summary metrics. **(B)** All fourteen *P. falciparum* chromosomes were assembled end-to-end, except for the left arm of chr. 1 which assembled as separate contig of 0.35 kb. The contig name in the SMRT assembly, its size, the corresponding chromosome and its coverage are indicated. The scale of coverage is between 0 (blue arrowhead) and 150-fold (red arrowhead) except for contig 64, where the scale is from 0-100. The centromere position in each contig is indicated with the green line. **(C)** Contig 12 of the SMRT assembly, which aligns to chr. 7, and its raw read coverage (black). **(D)** The depth of coverage of the ~4.5 kb PfCenH3-occupied region of chromosome 14, which averages 92% (A+T). The raw sequences obtained for a 176 bp fragment of the core centromere are shown. Colors: A = red, T = green, G = yellow and C = blue. **(E)** Depth of coverage of the ~34.6 kb telomeric/sub-telomeric region of chromosome 5. Gray reads are of high mapping quality, while red indicates a low mapping quality.

Identification of Large Genomic Variants

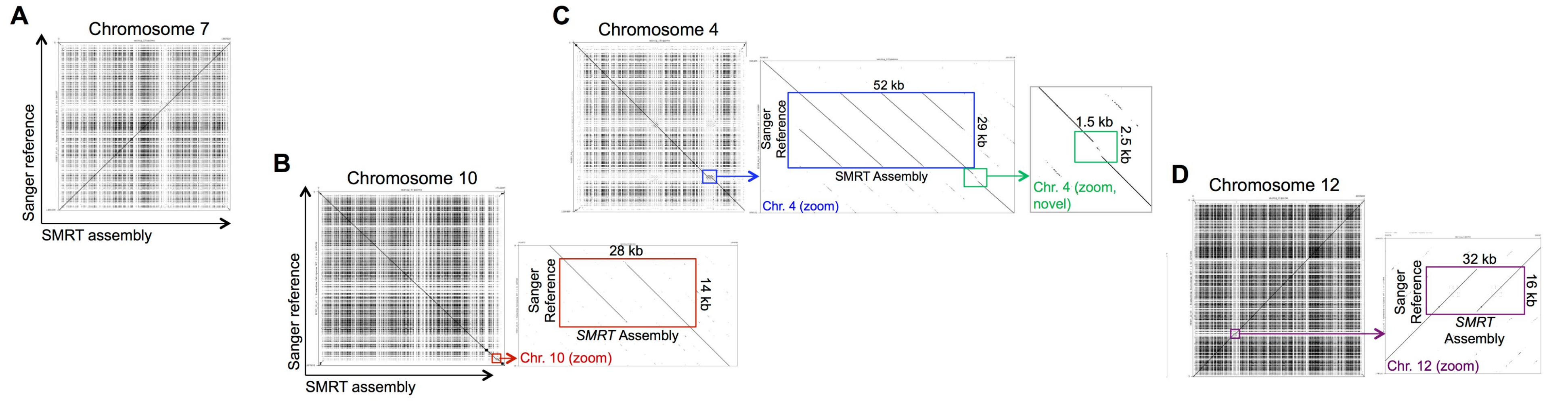
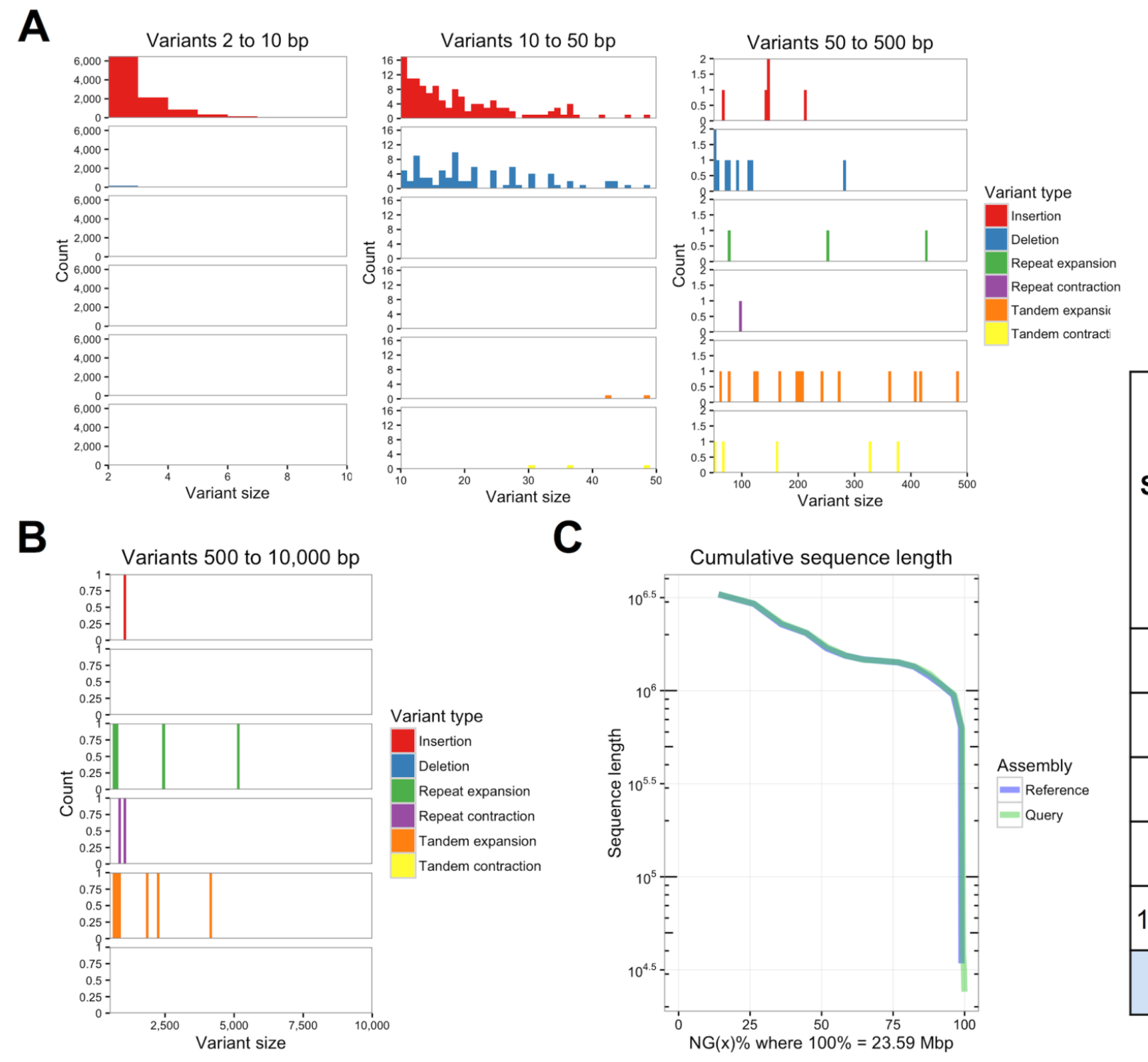


Figure 4. Dot plot comparisons of SMRT assembly to the reference 3D7 genome

Gepard[5] nucleotide alignments of *P. falciparum* SMRT assembly chromosomal contigs (X-axis) and chromosomes assembled from reference 3D7 genome (Y-axis; Sanger reference[6]). Each dot is a gray-scale representation of nucleotide identity within a 30-nucleotide window centered on that position. The main diagonal line shows the alignment between assemblies. **A)** Chromosome 7. **(B)** Chromosome 10. Zooming in to position 1,612,060, the SMRT assembly resolved a 14 kb tandem duplication relative to the reference. **(C)** Chromosome 4. Zooming in to position 939,044, the SMRT assembly resolved a 29 kb insertion. Further zooming to a position immediately downstream of this insertion shows a 1.5 kb stretch detected in the SMRT assembly with very low homology to the corresponding 2.5 kb stretch present in the reference genome, and is hence labeled 'novel'. **(D)** Chromosome 12. Zooming into position 1,707,528, the SMRT assembly resolved a 16 kb tandem duplication relative to the reference. Plots not drawn to scale.

Resolving New Structural Features



Size distribution of structural variants in the SMRT assembly relative to the reference genome

Size range (bp)	VARIANT TYPE														
	Insertion		Deletion		Tandem expansion		Tandem contraction		Repeat expansion		Repeat contraction			Homopolymer tract expansion (2-10 bp)	
	COUNT	TOTAL (bp)	COUNT	TOTAL (bp)	COUNT	TOTAL (bp)	COUNT	TOTAL (bp)	COUNT	TOTAL (bp)	COUNT	TOTAL (bp)		TYPE	COUNT
2-10	10112	26499	335	1148	0	0	0	0	1	8	0	0		poly-A	2535
10-50	130	2410	82	1786	2	90	3	114	0	0	0	0		poly-T	2517
50-100	1	66	6	401	2	142	2	123	1	77	1	97		poly-TA	1108
100-1000	4	647	3	507	15	5446	3	865	4	2123	1	808		poly-TG	3
1000-10000	1	1019	0	0	3	8294	0	0	2	7526	1	1018		poly-AG and poly-TC	1 each
Total	**10248**	**30641**	**426**	**3842**	**22**	**13972**	**8**	**1102**	**8**	**9734**	**3**	**1923**		**Total**	**6164**

Figure 5. Classification of structural variation types

Variants ranging from 2 bp to 10 kb in size were called using Assemblytics[7]. **(A)** Size distribution analysis of variants from 2 bp to 500 bp in size showed that the majority of insertions are <50 bp in size while the majority of deletions are between 10-500 bp. **(B)** Larger structural variants (500 bp to 10 kb) are depicted with the x-axis representing variant size in bp and the y-axis representing variant number. **(C)** Cumulative sequence length plot showing the nearly identical contiguity and total size of the SMRT assembly (query) versus the reference. The length of each individual sequence is indicated on the y-axis with the cumulative sum of sorted sequence lengths on the x-axis. The reference genome contig N50 (50% on the x-axis) is 1.688 Mb, while the SMRT Sequencing assembly contig N50 is 1.712 Mb.

Conclusions

With the high coverage of long reads, we were able to resolve the *P. falciparum* genome in unprecedented detail, despite its highly repetitive nature and high AT content. All but one chromosome was assembled telomere-to-telomere, and only one tandem repeat of >52 kb was not finished. The additional resolution to the repetitive elements contains previously hidden copies of genes vital to infectivity and virulence.