Article

Real time, field-deployable whole genome sequencing of malaria parasites using nanopore technology

Zahra Razook[1,2,3Ψ], Somya Mehra[1,2Ψ], Brittany Gilchrist[1,4], Digjaya Utama[1,4], Dulcie Lautu-Gumal[1,2,3,4,5], Abebe Fola[1,4], Didier Menard[6,7], James Kazura[8], Moses Laman[5], Ivo Mueller[1,4,6], Leanne J. Robinson[1,2,4,9], Melanie Bahlo[1,4] and Alyssa E. Barry[1,2,3,4*]

1. Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, AUSTRALIA
2. Life Sciences Discipline, Burnet Institute, Melbourne, AUSTRALIA
3. Institute for Mental and Physical Health and Clinical Translation (IMPACT), School of Medicine, Deakin University, Geelong, AUSTRALIA
4. Department of Medical Biology, University of Melbourne, Melbourne, AUSTRALIA
5. Vector Borne Diseases Unit, Papua New Guinea Institute of Medical Research, PAPUA NEW GUINEA
6. Institut Pasteur Paris, Paris, FRANCE
7. INSERM U1201, Paris, FRANCE
8. Case Western Reserve University, Cleveland, USA
9. Department of Epidemiology and Preventative Medicine, School of Public Health and Preventative Medicine, Monash University, Melbourne, AUSTRALIA

[Ψ]These authors contributed equally to the work

*Corresponding Author:
Institute for Mental and Physical Health and Clinical Translation (IMPACT), School of Medicine, Faculty of Health, Deakin University, 75 Pigdons Road, Waurn Ponds, Victoria, AUSTRALIA.
Ph: +61 3 52273504
E: a.barry@deakin.edu.au

ABSTRACT

Malaria parasite genomes have been generated predominantly using short read sequencing technology which can be slow, requires advanced laboratory training and does not adequately interrogate complex genomic regions that harbour important malaria virulence determinants. The portable Oxford Nanopore Technologies MinION platform generates long reads in real time and may overcome these limitations. We present compelling evidence that Nanopore sequencing delivers valuable additional information for malaria parasites with comparable data fidelity for single nucleotide variant (SNV) calls, compared to standard Illumina whole genome sequencing. We demonstrate this through sequencing of pure *Plasmodium falciparum* DNA, mock infections and natural isolates. Nanopore has low error rates for haploid SNV genotyping and identifies structural variants (SVs) not detected with short reads. Nanopore genomes are directly comparable to publically available genomes and produce high quality end to end chromosome assemblies. Nanopore sequencing will expedite genomic surveillance of malaria and provide new insights into parasite genome biology.

INTRODUCTION

Whole genome sequencing (WGS) provides complete information about pathogens that along with epidemiological metadata can enhance control efforts[1] to track the spread of infectious diseases in real time[2], the emergence of drug resistance[3], responses to control interventions[4] and to inform vaccine design[5]. Human malaria, a disease that has plagued humans for thousands of years, is caused by infection with *Plasmodium* species, the most virulent of which is *Plasmodium falciparum*. Malaria remains one of the world's most widespread and deadly infectious diseases killing more than 400 people annually, and causing over 200 million clinical episodes of the disease[6]. A major roadblock to WGS of clinical (or "field") isolates has recently been overcome through methods to enrich trace amounts of parasite DNA from finger prick blood samples contaminated with large amounts of human DNA[7, 8]. However, generation of WGS has traditionally been restricted to well-equipped laboratories that can maintain large, expensive sequencing platforms as well as advanced analytical pipelines and human resources required for data processing. As Illumina short read WGS (srWGS) requires extensive human expertise in both library preparation and instrument support, researchers in malaria-endemic countries have generally needed to send samples to large genome centres for sequencing resulting in significant time delays and loss of data custodianship. However for genomic surveillance to inform malaria control and elimination in a timely manner, large numbers of genomes with spatially dense sampling and the rapid generation of high quality data at low cost will be needed[9] and this is only feasible through sequencing in the field.

The Oxford Nanopore Technologies (ONT) MinION is a portable, pocket-sized device that operates through a USB port in a personal computer. Nanopore sequencing involves the movement of DNA through a synthetic porous membrane (flow cell) resulting in unique ionic currents for each six basepair window that are then translated into nucleotides in real time. The MinION device is an attractive option for malaria genomic surveillance because it can be deployed to field or clinic settings with minimal capital cost and training, and thus provides potential for rapid genomic profiling. In the clinical setting, MinION would provide a platform to rapidly screen for multigene haplotypes associated with drug resistance[10, 11], and may provide a pathway for personalised treatment in some well-resourced settings[12]. Long read WGS (lrWGS) using MinION could be used to rapidly track the origins of outbreaks as has been the case for other human pathogens[2, 13]. The low cost, relative portability and ease of use compared to other platforms suggests this platform has utility for both public health surveillance and capacity strengthening[14]. Data can be collected for future studies and can continue to be analysed for a variety of additional features as knowledge expands.

Unique features of the *Plasmodium falciparum* genome contribute to the challenge of WGS. The genome is AT-rich (81%), with extended tracts of repetitive, low-complexity DNA[15], particularly in subtelomeric regions. Hypervariable, multigene families, including the *var*, *rifin* and *stevor* families, have been difficult to characterise with srWGS because divergent short reads cannot be reliably aligned to a reference genome, although specialised pipelines have been developed[16]. In addition, malaria infections are often multiclonal, especially in high transmission regions[17] and therefore an added challenge for malaria population genomics is reconstructing these clonal haplotypes. LrWGS has the potential to overcome these limitations by spanning complex regions, to enable more accurate assembly and to differentiate clonal haplotypes within mutliple infections.

We aimed to develop and optimise Nanopore sequencing protocols for *P. falciparum* and to benchmark the resulting lrWGS against that obtained using the current gold standard Illumina srWGS. Initially, we tested the platform on pure *P. falciparum* DNA extracted from cultured strains, and compared the data to publically available reference genomes. We then optimised the sequencing protocol for natural human infections using mock infections comprising DNA from a reference *P. falciparum* strain spiked with human DNA. Finally, we performed shallow sequencing of a number of field isolates from Papua New Guinea (PNG) and Cambodia, allowing key parameters for field sample sequencing to be determined. The resulting data was used to estimate baseline error rates, to quantify the accuracy of single nucleotide variant (SNV) genotypes, to obtain drug resistance profiles, and to benchmark against publically available Illumina srWGS data from the same countries. Deep sequencing of cultured strains allowed *de novo* assemblies. These high quality reference genomes enable characterisation of full length *var* genes and mapping of large structural variants (SVs), which were not achievable with srWGS. Optimised sequencing protocols and pipelines are provided so that this approach can be implemented in other laboratories to permit true 'sequencing in the field'.

RESULTS

Nanopore sequencing of *P. falciparum*

In order to test the capability of MinION for sequencing *P. falciparum*, test runs were first performed using pure *P. falciparum* DNA from the XHA[18] and BB12 (a descendant of the Brazilian IT isolate[19, 20]) cultured lines with each run on an independent flow cell for the full 48 hours recommended by ONT. The total output of each run was 3.42 and 2.07 Gb, with N50 read lengths of 14.63 and 15.26 kB (mean read lengths of 8.80 and 6.95 kb) and maximum read lengths of 261.32 and 157.32 kb respectively (Table S2A), with quality scores of more than Q5 for over 90% of reads. The two samples produced 120X and 76X mean depth of coverage (reads), with 95% of the genome covered by more than 30 reads and relatively uniform genome-wide coverage (Figure 1A).

P. falciparum field isolates obtained by collecting the blood of human volunteers contain large amounts of contaminating human DNA after DNA extraction[21]. Mock infections of 1% parasitemia (20,000 parasites/μL[8]) were subject to different enrichment conditions. The resulting six samples were run in multiplex on a single 48hr run. The enrichment procedure (*Mcr*BC digest plus random whole genome amplification (rWGA)) significantly increased the breadth and depth of parasite genome coverage (Figure 1B, Table S2B) and the proportion of parasite DNA (≥83%) compared to rWGA alone (≤5%). After end repair, the 0.45V bead clean-up was associated with higher genome-wide read coverage than 1V bead clean-up. However, both 2.5V ethanol precipitation and 1.8V SPRISelect bead purification prior to library preparation performed similarly in terms of the read quality and output (Table S2B).

Using the optimised Nanopore sequencing protocol (1.8V bead purification after enrichment and 0.45V beads purification after end prep), we then sequenced 33 *P. falciparum* field isolates, multiplexing three to four samples per 48 hour run (Table S2C). Field samples ranged in starting parasite densities from 0.674–32,200 18s rRNA copies/uL (Q1=916 copies/uL, Median=3,000 copies/uL, Q3=16,840 copies/uL). Methylation-dependent *Mcr*BC digestion and whole genome amplification resulted in between 114–810,000 fold enrichment (Q1=7,940, Median=21,200, Q3=52,900). The proportion of reads mapping to the parasite genome varied between 0% and 85% (Q1=6%, Median=18%, Q3=36%). Median read lengths ranged from 990–2,790 bp.

Key parameters associated with sequencing quality for individual field isolates are shown in Figure 1C-D. After enrichment, parasite DNA concentration (as measured by 18s rRNA copy number per μL) exhibited the strongest correlation with the breadth of genome wide coverage at 1X (R=0.77, p=1.2E-7) (Figure 1C). A copy number threshold of 1.0×10^8 was found to be the most appropriate diagnostic for ≥75% breadth of coverage at 1X in a receiver-operating characteristic curve analysis (data not shown). Based on these results, enriched samples with a minimum of 1.0×10^8 copies/μL are recommended for successful sequencing. A comparison of copy numbers before and after enrichment (Figure 1D) suggests that DNA samples with a minimum of 1000 copies/μL should be selected for enrichment.

Stratification by flow cell showed an association between the sequencing output and the amount of DNA loaded into a flowcell (Figure S2). A general, decreasing trend between the total number of bases sequenced and the amount of DNA loaded was evident above 600ng. Similarly, the proportion of reads with quality scores above 10 decreased as the amount of DNA loaded into a flowcell increased. Hence, clogged pores, were more likely when higher amounts of DNA (>600 ng) were loaded into the flowcell.

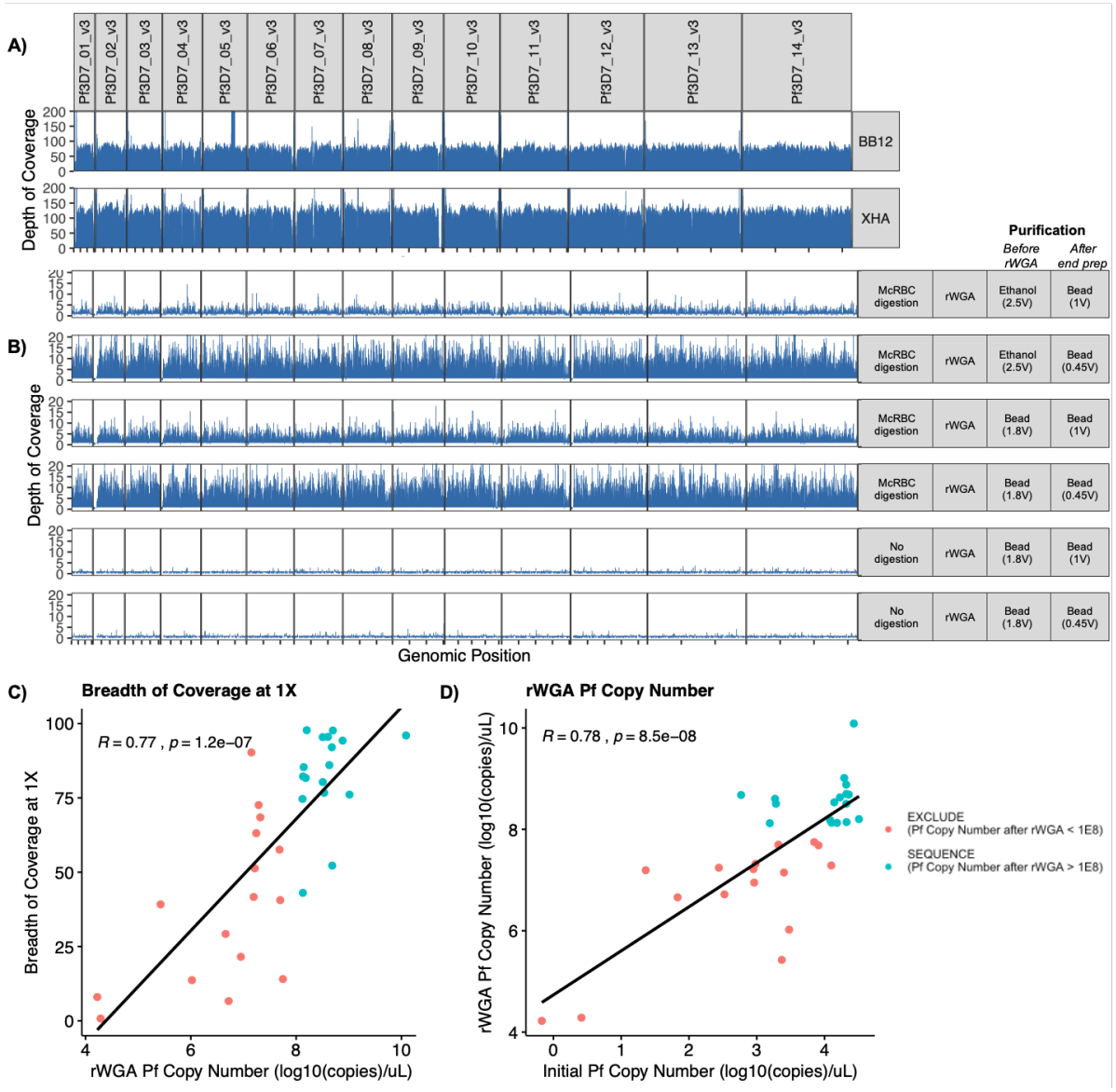


Figure 1. Output of MinION sequencing runs. (A) *Schematic of genome coverage assessed in 5kb bins of the parasite genome for pure P. falciparum clones (cultured lines). Coverage is a term that can be used interchangeably to describe both depth and breadth of genome coverage. Depth refers to the number of times a base in the reference genome is covered by a read and correlates to greater confidence in variant calling. Breadth refers to the percentage of the target genome that is covered by reads in a sequencing experiment. High uniformity, breadth and depth of coverage were attained for both BB12 and XHA.* ***(B)*** *Optimisation of parasite DNA amplification and library preparation using mock infections. Samples comprising 1% P. falciparum DNA (3D7) were spiked with human DNA and subject to different enrichment (digestion with McrBC and/or rWGA). The depth of sequencing coverage across the parasite genome is shown for each condition. Sequencing was run in multiplex for the six conditions.* ***(C)*** *Log-transformed post-enrichment P. falciparum 18s rRNA gene copy numbers (as a measure of Pf DNA concentration within a sample) are most strongly correlated with the breadth of coverage at 1X (the proportion of the genome covered by at least one read). A Pf copy number threshold of with 1.0×10^8 is the most appropriate diagnostic for a minimum breadth of coverage of 75% at 1X based on a receiver-operator characteristic analysis (data not shown).* ***(D)*** *rWGA Pf copy numbers after enrichment as a function of the pre-enrichment (initial) Pf copy number. Datapoints are coloured on the basis of whether they met the criteria for exclusion (red) or sequencing (green). Based on these findings we recommend selecting field isolates with a minimum of 1000 parasite copies/µL for enrichment, and enriched samples with at least 1.0×10^8 parasite copies/µL for sequencing.*

Baseline sequencing error rates in Nanopore and Illumina sequencing

To quantify baseline error rates for Nanopore and Illumina, we compared 3D7 WGS to the publicly available *P. falciparum* 3D7 reference genome (version 3). While the core *P. falciparum* genome is stable in long-term *in vitro* culture, mitotic recombination can drive variation in antigen genes and subtelomeric regions[22]. We therefore restricted the reference genome-based analyses to SNVs situated in the core genome, with reference alleles to be treated as truth calls and alternate alleles as error. We focused on two key use cases for SNV calling: *de novo* SNV discovery (e.g. to interrogate novel variants associated with drug resistance) and SNV genotyping at a set of known loci (e.g. to conduct population genomic analyses or drug resistance profiling).

First, we considered false discovery rates (FDRs) for the detection of haploid *de novo* SNVs using each platform focusing on coding SNVs in 1356 "housekeeping" genes essential for *in vitro* asexual blood stage development[23], with a cumulative transcript length of 2,792,980 bp. Illumina detected no coding SNVs in these genes relative to the reference genome while Nanopore sequencing data gave rise to 15 coding SNVs after relatively lenient filtration (i.e. minimum depth of coverage 5X, with at least 75% of reads supporting the called allele), with seven SNVs retained when the minimum depth of coverage was increased to 10X. These results suggest a higher FDR for *de novo* SNV discovery using Nanopore data relative to Illumina data, though it remains almost negligibly low. Albeit low coverage will also lead to lower power to detect novel variants. FDRs are contingent on coverage and variant filtration parameters - unlike the Illumina-specific GATK pipeline which considers well-validated metrics to filter out poor alignments, best practices workflows are yet to be developed for Nanopore SNV genotyping.

Next, baseline error rates for Nanopore and Illumina genotypes at a set of 742,365 validated SNV loci were analysed. Of 684,737 SNVs successfully genotyped by both platforms, Illumina incorrectly called 13 alternate alleles, while Nanopore incorrectly called 40 alternate alleles after minimal filtration (i.e. minimum depth of coverage 2X, with at least 75% of reads supporting the called allele). Baseline error rates for genotyping SNVs at validated loci therefore seem comparable for Nanopore and Illumina and appear to be very low across both platforms (<0.006%). However, since we have mapped reads against an isogenic reference genome, read alignments are likely to be correct and high genotyping accuracy may be driven by reference bias. It is therefore important to perform additional benchmarking of Nanopore and Illumina genotypes for isolates for which there is no isogenic reference genome.

Accuracy of Nanopore genotypes

WGS analyses of *P. falciparum* commonly utilize high quality SNV calls to conduct population genomic analyses[24]. To validate the genotyping accuracy of Nanopore and analysis pipeline relative to Illumina, we performed a comparison of SNV genotypes derived from strains BB12 (Brazil) and XHA (Papua New Guinea, PNG) which have very high coverage (>70X), and two Cambodian field isolates with low coverage (<5X) for which we also collected Illumina data. Data was aligned to the 3D7 reference genome (version 3) and variants called as per 3D7 analysis. We then considered allele calls at 742,635 high-quality variants obtained through an in-house re-analysis of the srWGS MalariaGEN Pf3k plus PNG dataset[24, 25] comprising 2,661 field isolates sampled from 15 malaria endemic countries. Only haploid genotypes determined by calling a single dominant allele at each locus were considered.

Nanopore SNV calls exhibited strong bias towards reference allele calls at very low depths of coverage. Moreover, at low depth, alternate (non-reference) alleles were generally called as reference alleles by the Nanopore analysis pipeline when the proportion of reads supporting the called genotype was relatively low, suggestive of conservative allele calls. Reference

alignment bias, which arises during read mapping, has previously been reported for Nanopore (MinION) SNV genotyping[26], with implications for haplotype reconstruction due to the overrepresentation of reference haplotypes[27].

Since Nanopore sequencing has been shown to exhibit significant systematic error in homopolymeric regions[28] we also stratified allele calls based on genomic content, by differentiating between calls within homopolymeric tracts of length greater than 6 bp and those in non-homopolymeric regions. Alignments around discordant calls in homopolymeric tracts were poor, with lower coverage than flanking genomic regions and a smaller proportion of reads supporting the called allele (Figure 2). In non-homopolymeric regions, discordant calls exhibited similar signatures, with a lower proportion of reads mapping to the called allele, but less substantial reductions in coverage. Discordant calls between Nanopore and Illumina could thus be diagnosed on the basis of low coverage and a low proportion of reads mapping to the called allele (that is, false heterozygosity) (Figure 2). Poor alignments of Nanopore sequence data around homopolymer tracts[28] and false heterozygosity due to Nanopore sequencing error, particularly at low depths of coverage[29], have been previously documented.

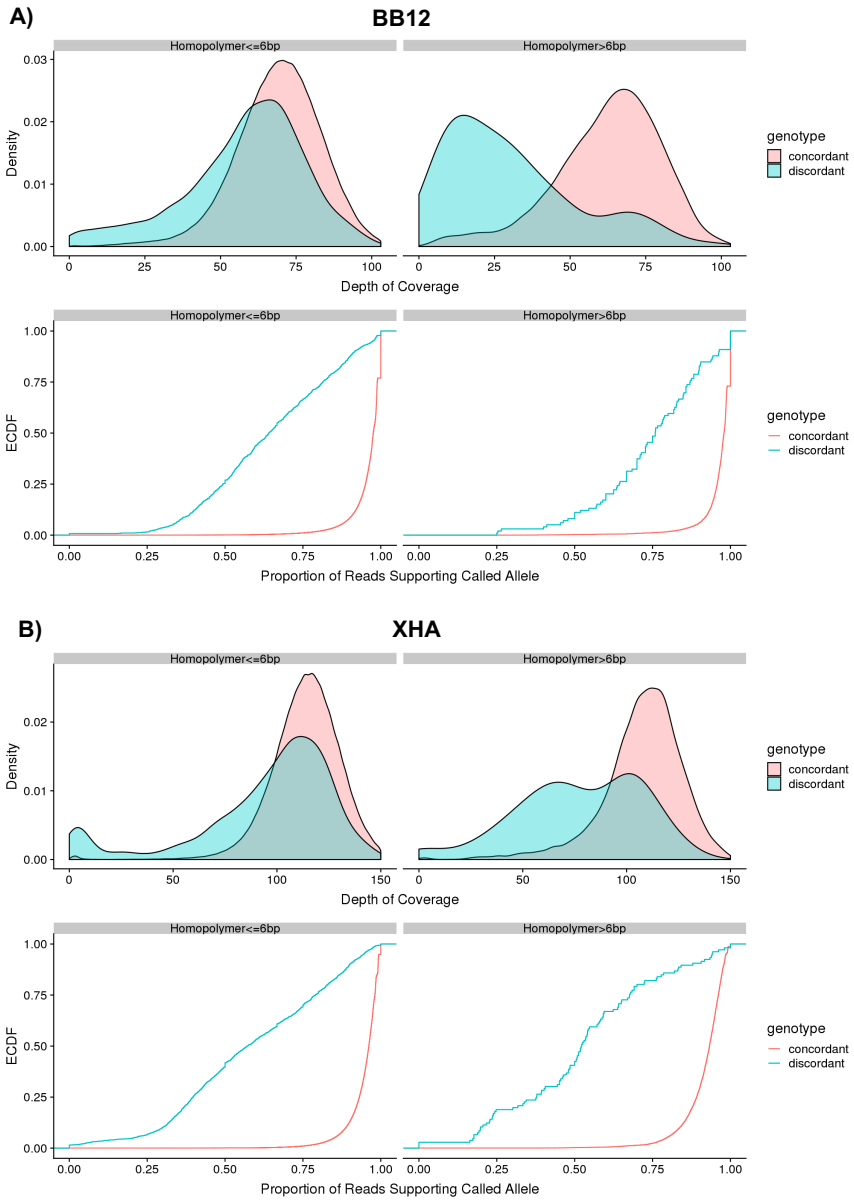


Figure 2. Variant quality metrics for discordant and concordant MinION genotype calls in both homopolymeric (>6bp) and non-homopolymeric regions *for (A) BB12 and (B) XHA. Density plots*

show distributions for the depth of coverage by locus. Empirical cumulative distribution functions (ECDFs) show distributions for the proportion of reads supporting the called allele. In both homopolymeric and non-homopolymeric regions, discordant calls frequently had lower depths of coverage and a smaller proportion of reads supporting the called allele than concordant calls.

As a strategy to remove artefacts, loci were filtered by the depth of coverage and the proportion of reads supporting the allele call; previous studies have adopted these metrics[10]. Excluding loci with read support across multiple alleles does not account for heterozygosity or multiclonality. However, this approach was deemed appropriate since the cultured strains were presumed to be monoclonal, and shallow sequencing of field isolates generated inadequate coverage to detect minor clones. For field isolates (Table S2C) we only retained those genotypes with a minimum depth of coverage 2X where at least 75% of reads supported the called genotype (i.e. if 2X coverage, both reads must support the call).

Table 1: Comparison of Haploid MinION and Illumina Genotypes at 742,365 High-Quality Loci

Sample: XHA (106X coverage)

	Illumina Call	*MinION Call*	*Unfiltered*	*Filtered*
Concordant	REF	REF	683,648	677,524
	ALT	ALT	7,906	6,963
Discordant	REF	ALT	612	336
	ALT	REF	1,133	180
	ALT 1	ALT 2	10	10
Missing	N/A	N/A	49,056	57,352
Percentage of correct MinION REF calls:			99.83	99.97
Percentage of correct MinION ALT calls:			92.71	95.27
Overall concordance rate:			99.75	99.92

Sample: BB12 (75X coverage)

	Illumina Call	*MinION Call*	*Unfiltered*	*Filtered*
Concordant	REF	REF	656,151	650,085
	ALT	ALT	8,301	7,335
Discordant	REF	ALT	489	224
	ALT	REF	776	217
	ALT 1	ALT 2	13	12
Missing	N/A	N/A	76,635	84,492
Percentage of correct MinION REF calls:			99.88	99.97
Percentage of correct MinION ALT calls:			94.30	96.88
Overall concordance rate:			99.81	99.93

Sample: Cam_01 (3.2X coverage)

	Illumina Call	*MinION Call*	*Unfiltered*	*Filtered*
Concordant	REF	REF	453,740	304,211
	ALT	ALT	3,156	2,518
Discordant	REF	ALT	1,243	322
	ALT	REF	2,412	208
	ALT 1	ALT 2	34	26
Missing	N/A	N/A	281,780	435,080
Percentage of correct MinION REF calls:			99.47%	99.93%
Percentage of correct MinION ALT calls:			71.19%	87.86%
Overall concordance rate:			99.20%	99.82%

Sample: Cam_02 (1.4X coverage)

	Illumina Call	*MinION Call*	*Unfiltered*	*Filtered*
Concordant	REF	REF	331,371	170,256
	ALT	ALT	1,831	1,296
Discordant	REF	ALT	981	171
	ALT	REF	2,225	122
	ALT 1	ALT 2	23	12
Missing	N/A	N/A	405,934	570,508
Percentage of correct MinION REF calls:			99.33%	99.93%
Percentage of correct MinION ALT calls:			64.59%	87.63%
Overall concordance rate:			99.04%	99.82%

Overall discordance rates between haploid Nanopore and Illumina genotypes were consistently low, even for Cambodian field isolates with poor mean coverage (less than 4X) however these isolates also had a large proportion of missing loci after filtering (Table 1). Nanopore reference calls had accuracy above 99.9% for all four isolates (Table 1). However, alternate allele calls were more error prone even after filtration, with only 88% supported by Illumina for Cam_01 (3.2X coverage) and Cam_02 (1.4X coverage). For cultured lines BB12 (75X coverage) and XHA (106X coverage) however, over 95% of Nanopore alternate allele calls were supported by Illumina calls. These results indicate that Nanopore alternate allele calls are significantly less reliable at low depths of coverage. As expected, discordance rates in homopolymeric tracts greater than 6 bp were found to be substantially higher than overall discordance rates (Table S3), highlighting the need to exercise caution in analysis of calls from homopolymeric tracts.

Random errors in Nanopore sequencing reads were mitigated by increasing coverage, as evidenced by the lower discordance rate for isolates with high coverage. However, the

persistence of genotyping error at high depths of coverage points to systematic differences between the two platforms. We therefore attempted to identify characteristic features of concordant and discordant calls that passed quality filtration.

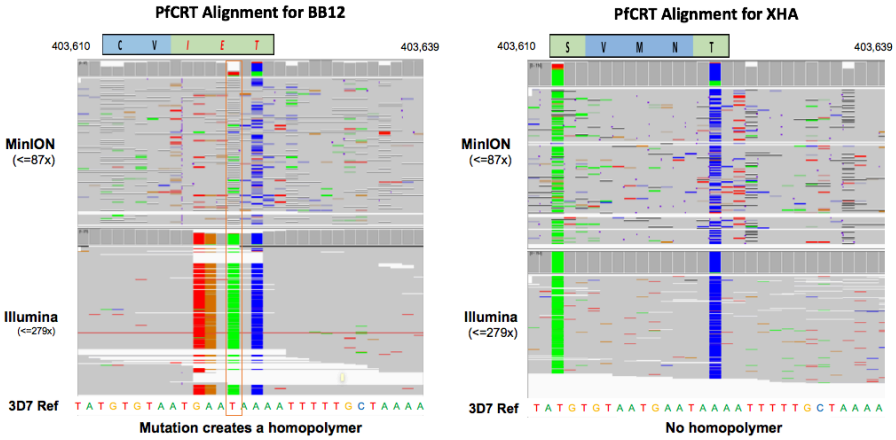
The relative frequencies of concordant and discordant genotype calls after filtration were stratified by the variation type at discordant sites found within Nanopore and Illumina data. The most common discordant genotypes across the trialled isolates were the Illumina (treated to be truth calls) ⟶ Nanopore (treated as erroneous calls) transitions G ⟶ A and C ⟶ T, followed by the transversions T ⟶ A and A ⟶ T (Figure S3). In general, transition errors between similarly-structured bases (i.e. between two-ring purines or one-ring pyrmidines) were more pronounced than transversion errors for Nanopore sequencing, which might reflect errors in the base calling algorithm. Given that Nanopore calls are generated based on minute changes in current as the DNA passes through the pore, it is not surprising that biochemically similar bases are harder to distinguish and result in erroneous calls.

Drug resistance profiling using Nanopore sequencing

As a measure of the capability of Nanopore sequencing to produce functionally relevant information from skim sequencing of field isolates, we determined SNV genotypes for each field isolate across a range of known drug resistance marker loci including *crt* (chloroquine, amodiaquine, piperaquine), *dhfr-ts* (pyrimethamine), *dhps* (sulfadoxine), *mdr1* (lumefantrine, mefloquine) and *kelch13* (artemisinin). Only genotype calls with depth of coverage at least 2X and at least 75% of reads supporting the called allele were retained. Functional annotations were applied to filtered variants, allowing the extraction of both nucleotide and amino acid haplotypes.

Haplotypes generated from Nanopore and Illumina sequencing data for the cultured lines BB12 and XHA, and field isolate Cam_01 were compared first. Filtered drug resistance gene haplotypes were generally concordant between the two sequencing platforms, with the exception of CRT codons 74 to 76. Here, for isolates BB12 and Cam_01, alternate alleles were incorrectly called as reference alleles using Nanopore data (Figure 3). For Cam_01, reference alignment bias due to low coverage may have contributed to the erroneous genotype call at CRT codon 74; BB12, however, had high overall coverage. Inspection of the alignments in this region revealed that for BB12, one of the resistance-associated variants gave rise to a homopolymer. No such homopolymer was introduced by the XHA mutation (Figure 3A), hence resulting in concordant haplotypes. Since Nanopore sequencing is subject to systematic error in homopolymeric stretches, this again demonstrates that caution should be exercised when potential homopolymeric tracts are encountered in loci selected for genotyping. However, we note that recent advances in Nanopore flow cell chemistry and improved basecalling algorithms will help mitigate these errors further.

A)



B)

Origin	Year	Sample ID		Gene	DHFR-TS				MDR1			CRT									DHPS			K13				
				Codon	16	51	59	108	86	184	1042	72	74	75	76	220	271	326	356	371	436	437	540	493	539	543	553	580
					Pf4.748134	Pf4.748239	Pf4.748262	Pf4.748410	Pf5.958145	Pf5.958440	Pf5.961013	Pf7.403612	Pf7.403613	Pf7.403620	Pf7.403621	Pf7.403623	Pf7.403625	Pf7.404407	Pf7.404836	Pf7.405361	Pf7.405382	Pf7.405599	Pf7.405600	Pf7.405838	Pf8549681	Pf8549682	Pf8549685	Pf8549993	Pf13.1725521	Pf13.1725382	Pf13.1725370	Pf13.1725340	Pf13.1725259
CULTURED LINE	IT4 descendant	–	BB12	(Illumina)	V	N	C	T	Y	Y	N	C	I	Q	T	S	E	S	I	I	S	A	K	Y	R	I	P	C
				(MinION)									M	N	–													
	Papua New Guinea	–	XHA	(Illumina)	A	N	C	S	Y	Y	N	S	M	N	T	S	Q	D	L	R	S	A	K	Y	R	I	P	C
				(MinION)																								
FIELD ISOLATE	Cambodia	?	Cam_01	(Illumina)	A	I	R	N	N	Y	N	C	I	Q	T	S	E	N	I	I	S	A	K	Y	R	I	P	C
				(MinION)	–	–							M	K			–		–	–				–			–	
	East Sepik, Papua New Guinea	2005	PNG_05_1		A	N	C	S	–	Y	N	–	M	N	T	–	Q	D	L	–	–	–	–	Y	R	–	–	C
			PNG_05_2		A	N	–	N	–	Y	N	C	M	N	K	A	Q	N	I	R	S	–	K	Y	R	I	P	C
			PNG_05_3		A	N	–	N	Y	Y	N	S	M	N	T	A	Q	–	L	R	S	–	K	Y	R	I	P	C
			PNG_05_4		–	–	–	N	–	Y	N	–	–	–	–	–	Q	D	–	–	–	–	–	–	R	I	P	C
			PNG_05_5		A	N	–	N		Y	N	S	M	N	T	S	Q	D	L	R	S	–	K	Y	R	I	P	C
			PNG_05_6		A	N	–	–	Y	Y	N					–			L	R	S	–	K	Y	R	I	P	C
			PNG_05_7		A	N	–	N	Y	Y	N	–	M	N	–	S		D	L	R	S	–	K	Y	R	I	P	C
			PNG_05_8		A						N	–	–	N	K	A	Q	N	I	R	C	A	K	Y	R	I	P	C
		2012	PNG_12_1		A	N	R	N	N	Y	N	S	M	N	T	S	Q	–	L	R	S	G	E	Y	R	I	P	C
		2016	PNG_16_1		A	N	R	N	N	F	N	S	M	N	T	S	Q	D	L	R	S	G	E	Y	R	I	P	C
			PNG_16_2		A	N	R	N	N	Y	N	S	M	N	T	S	Q	D	L	R	S	G	E	Y	R	I	P	C
			PNG_16_3		A	N	–	N			N	–	M	N	–	–	–	–	L	R	–	–	E	Y	R	I	P	C
			PNG_16_4						N	–		–	M	N	–	–					S		K	Y	R	I	–	C
			PNG_16_5		–	N		N	N	Y		S	M	N	–	–	Q	–	–	R	S	G	K	Y	R	I	P	C
			PNG_16_6		–	N	–	N		Y	N	S	M	N	–	S	Q	D	L	R	S	A	K	Y	R	I	P	C
			PNG_16_7		A	N		N			D	S	M	N	–	S	Q	D	L	R	S	A	K	Y	R	I	P	C
			PNG_16_8		A	N	R	N	N	F	D	–	M	N	–	–	–	–	L	R	S	G	E	Y	R	I	P	C

LEGEND
- wildtype (reference allele)
- mutant (alternate allele)
- – low quality genotype (filtered)
- no genotype (missing data)

Figure 3*: Drug resistance profiling of MinION field isolates. (A) Alignments of MinION sequence data across codons 72 to 76 of CRT for cultured lines BB12 and XHA. MinION and Illumina genotypes are discordant for BB12, where the mutation at position Pf3D7_07_v3:406325 has given rise to a homopolymeric tract. MinION and Illumina genotypes are concordant for XHA, for which there is no such homopolymeric stretch. (B) Drug resistance haplotypes for field isolates with breadth of coverage >65% at 1X. Isolates were genotyped at loci corresponding to key resistance-associated variants in genes crt, mdr1, dhfr, dhps and k13. Crt variants previously associated with resistance that were screened included T93S, H97Y, F145I, M343I and G353V however all isolates had wildtype alleles. Red boxes indicate resistance alleles, while wildtype alleles are indicated in blue. Missing genotypes (for which there was no sequence data) and low quality genotypes (which were removed after quality filtration) are shown in white and grey respectively.*

Amino acid haplotypes spanning key resistance-associated variants across drug resistance-associated genes are presented in Figure 3B. Only Nanopore sequenced field isolates with a minimum breadth of coverage 65% at 1X are presented, and except for *kelch13*, only polymorphic loci within the dataset are shown. Some genotype calls were filtered due to low coverage. Missing genotype calls (occurring when no reads align to a particular locus) tend to be clustered, since genome-wide coverage for some Nanopore field isolates was uneven with no data captured for some genomic regions. These issues should be resolved as methods for parasite genomic DNA enrichment improve.

Nanopore whole genome sequences are directly comparable to publically available P. falciparum Illumina data

Given that the majority of publically available *P. falciparum* genomes have been determined using Illumina short read sequencing, we asked whether genotypes obtained using Nanopore sequencing could be directly compared with these genotypes or whether Nanopore sequencing gave rise to platform-specific effects, thus preventing any biological insight from direct comparisons. We examined clustering patterns for Nanopore-sequenced field isolates (originating from PNG or Cambodia) relative to Illumina-sequenced isolates (originating from PNG, Cambodia, Vietnam, Laos or Thailand) from the MalariaGEN Pf3k/PNG dataset.

Each isolate was genotyped at a set of 742,365 high-quality SNVs identified thrugh an in-house reanalysis of the MalariaGEN P3k/PNG dataset. Genotype missingness filtration by isolate (<30%) and SNV loci (<10%) was performed to avoid possible biases introduced by imputation. Monomorphic loci were removed since they were uninformative, but all polymorphic SNVs were retained; SNVs were also not filtered by minor allele frequency to avoid ascertainment bias. After filtration, we retained 10 Nanopore- and 1040 Illumina-sequenced field isolates, genotyped at 51,421 polymorphic SNV loci.

Principal coordinates analysis (PCoA) was then performed on pairwise genetic distances between isolates, defined to be the proportion of successfully-genotyped loci with shared alleles for each pair of isolates. Projections of isolates onto two dimensions are visualised in Figure 4. Expected geographical clustering patterns emerged, with field isolates originating from PNG and South East Asia generally clustering distinctly. Nanopore- and Illumina-sequenced field isolates originating from PNG cluster together closely, suggesting that the different sequencing platforms do not impact the clustering analysis, with biological signals dominating instead.

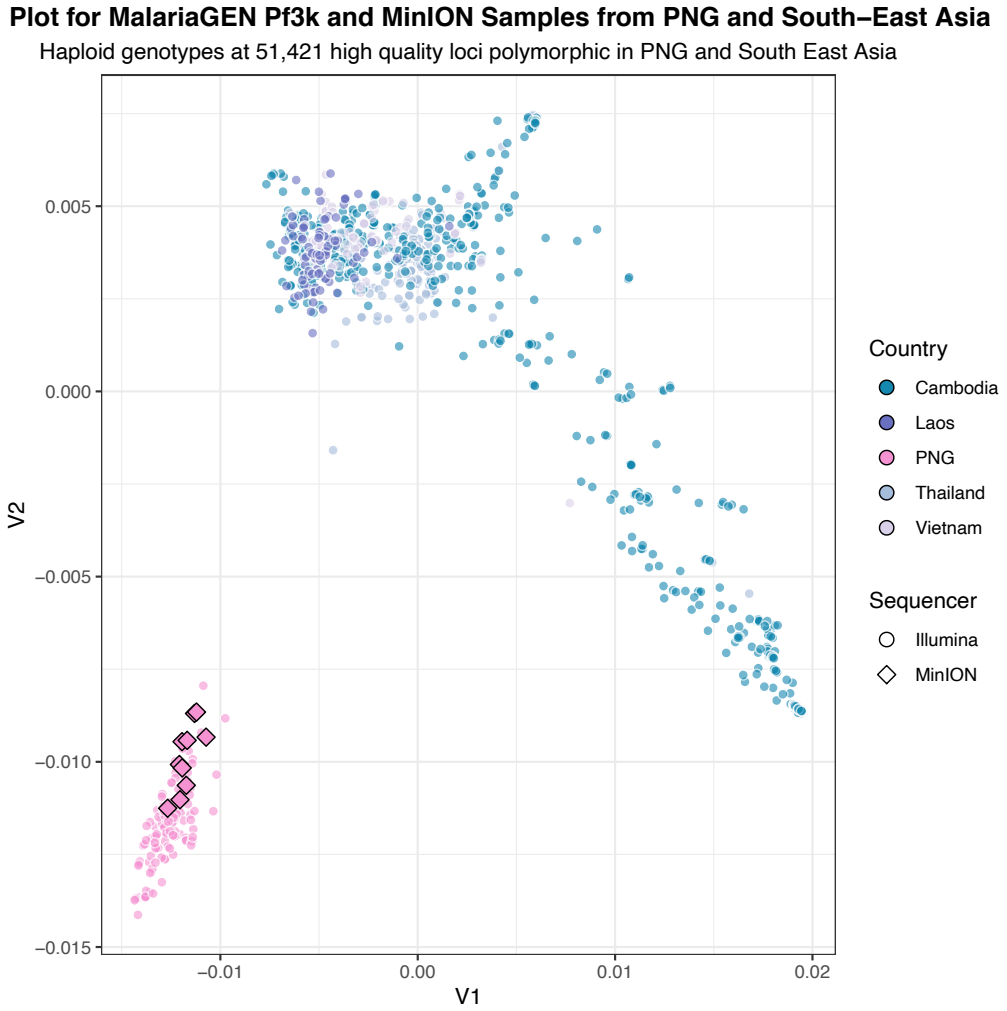


Figure 4: Clustering of Nanopore and Illumina-sequenced field isolate genomes from Papua New Guinea and South-East Asia (Cambodia, Laos, Thailand and Vietnam). *Whole genome sequence data was used to genotype 742,365 high-quality SNVs obtained from an in-house reanalysis of the MalariaGEN Pf3k/PNG dataset. Ten Nanopore and 1,040 Illumina-sequenced field isolates were retained after samples with genotype missingness above 30% were filtered. Filtration of SNPs with missingness above 10% yielded 51,421 polymorphic SNVs. Pairwise distances between isolates, based on the proportion of the genotype shared between pairs, were calculated using the filtered set of SNVs. PCoA was performed on the resultant distance matrix to obtain a two-dimensional representation of the data. Each point represents a distinct isolate, with colours representing geographic origins, and shapes representing sequencing platforms. Nanopore- and Illumina-sequenced field isolates originating from PNG cluster together closely in the PCoA.*

De novo genome assembly using Nanopore long reads allows characterisation of highly polymorphic regions

De novo genome assembly can further improve characterisation of parasite genomes, enabling the inclusion of highly-polymorphic genomic regions that are typically blacklisted during variant calling from srWGS due to extreme polymorphism relative to the reference genome. The advent of lrWGS has greatly enchanced the continuity and completeness of *de novo* genome assembly, particularly in low-complexity and repetitive genomic regions that have been difficult to resolve with short read fragments (<1000bp)[30].

To characterise repetitive and highly polymorphic genomic regions, we generated *de novo* genome assemblies using a combination of long and short read data. Long, continuous scaffolds were first generated using Nanopore long read data. Individual base errors, indels, block substitution events, gaps and local misassemblies in the scaffolds were then corrected

using Illumina short read data, an approach that has also been used by others (https://github.com/nanoporetech/ont-assembly-polish). Summary statistics for the polished *de novo* assemblies are presented in Table 2.

Table 2: Summary statistics for *de novo* assemblies of *P. falciparum* genomes

	BB12 (IT ref)	XHA (3D7 ref)
No. contigs	17	15
Total length (bp)	23,097,724	23,117,150
Longest contig (bp)	3,350,589	3,384,513
N50 length (bp)	1,595,563	1,599,767
N75 length (bp)	1,375,631	1,353,651
Shortest contig (bp)	64,574	230,110
L50 (no. contigs)	5	5
L75 (no. contigs)	9	9
GC content (%)	19.47	19.23

The BB12 assembly was compared against the reference genome of its closest relative, the parent IT4 strain (version 4). However, for XHA, which is an isolate from PNG, no close relative reference genome is available so we compared the *de novo* assembly against the 3D7 reference genome; the 3D7 reference contains 14 complete nuclear chromosomes and has higher continuity than the IT4 reference genome. Assembly continuity for both BB12 and XHA was high, with all nuclear chromosomes except chromosome 5 spanned by a single contig (Figure S4). Breakpoints for chromosome 5, which was spanned by two contigs in both assemblies, varied for XHA and BB12. Alignment of the BB12 assembly against the IT4 reference genome (version 4) revealed that the two shortest contigs were spurious, multi-mapping to the subtelomeric regions of several nuclear chromosomes in addition to several contig fragments.

Although *de novo* contigs and reference chromosomes were generally concordant in core genomic regions, alignments in subtelomeric hypervariable regions were more fragmented. This fragmentation could reflect true variation between cultured lines and reference isolates (e.g. XHA vs 3D7), but may also be a marker of translocation and contig misassembly for homologous lines (e.g. BB12 vs IT4). A 150 kbp translocation from chromosome 13, for instance, is apparent in the downstream subtelomeric region of contig PfBB12_11.

The BB12 assembly allowed a more complete characterisation of chromosomes 6 and 13 relative to the IT4 reference genome (version 4). Since our BB12 assembly exhibited higher continuity than the IT4 assembly (version 4), we were able to localise fragment PfIT_00_11 to the subteolomeric region of chromosome 6, while fragments PfIT_00_4 and PfIT_00_10 were identified to be neighbouring subtelomeric flanks of chromosome PfIT_13 (Figure 5).

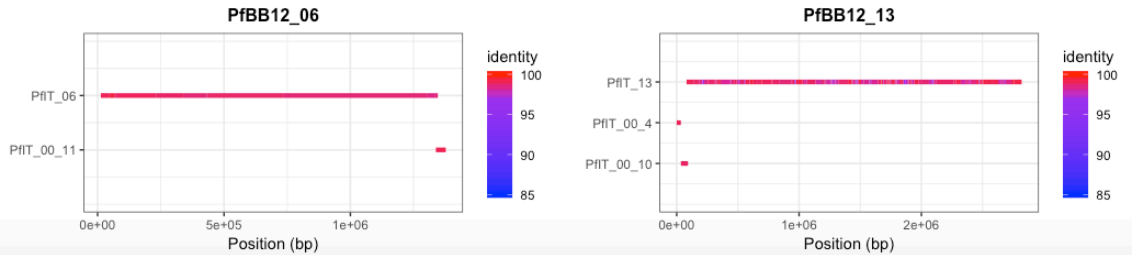


Figure 5: Enhanced continuity of MinION BB12 assembly, relative to the IT4 reference genome (version 4). *More complete assembly of chromosomes 6 and 13 was enabled through long read sequencing. Fragment PfIT_00_11 has been localised to the subtelomeric region of PfIT_06, while fragments PfIT_00_4 and PfIT_00_10 have been identified as neighbouring subtelomeric regions of chromosome 12. One-to-one mappings between de novo contigs and reference genomes have been computed using nucmer (MUMmer, V.3.1[31]).*

A genome-wide summary of annotated features in our assemblies for XHA and BB12 is shown in Table 3, with corresponding statistics for the 3D7 reference genome shown, as a comparison. The elevated number of pseudogenes identified relative to the 3D7 reference genome, in addition to the lower annotated gene density, suggests that open reading frames (ORFs) in the *de novo* assemblies were disrupted by mismatch and indel errors in a number of instances. Further, some annotated genes were fragmented, with multiple neighbouring annotations likely corresponding to the same gene. However, we were still able to characterise a large number of genomic features in our assemblies.

Table 3: Summary of annotated genomic features for *de novo* assemblies of *P. falciparum* genomes compared to the 3D7 reference genome.

	BB12 assembly	XHA assembly	3D7 reference
No. genes	4,762	4,491	5,540*
Gene density (genes/Mb)	203.4	191.46	237.8*
No. coding genes	4,698	4,426	5,362*
No. non-coding genes	64	65	178*
No. pseudogenes	1,081	1,491	153*
Coding GC%	24.7	24.95	23.7[15]

* Ensembl protist database

We identified 27 complete, 16 split and 11 partial *var* genes for BB12, and 18 complete, 16 split and 15 partial *var* genes for XHA. Domain structures for all complete and partial *var* genes are shown in Figure S5.

Improved structural variant calling using long read data

Approaches utilising srWGS tend to provide limited resolution for the detection of structural variants (SVs)[32]. LrWGS has emerged as a promising tool for SV detection, particularly complex SV. Concordance rates across three trialled SV pipelines are presented in Figures 6A and 6B, focussing on SVs longer than 200bp to avoid systematic errors associated with Nanopore sequencing[33]. SVs detected using long reads only (Sniffles[33]) and our *de novo* assemblies (Assemblytics[34]) exhibited reasonable levels of concordance, with approximately 30% of high-quality SVs consistent across both pipelines. SVs detected using short reads only (GRIDSS[35]) had very little concordance with either Sniffles or Assemblytics called SVs. Short read SV calling (GRIDSS) identified fewer high-quality SVs longer than 200 bp, with the vast majority of high-quality SVs within the length range 30-200 bp. Longer SVs failed to pass default quality filtration parameters. Distributions of SV types and lengths for the three pipelines are shown in Figures 6C and 6D.

We then screened high-quality SVs to identify functionally-relevant variation. For BB12, a 96kbp duplication on chromosome 5 spanning *mdr1 (*PF3D7_0523000) was identified by both Sniffles and Assemblytics, but not GRIDSS. A quantitative real-time PCR (qPCR) assay confirmed the presence of the duplication spanning *mdr1* (data not shown). The amplification of *mdr1* has been associated with resistance to a range of antimalarial drugs, including artemisinin, lumenfantrine, mefloquine and halofantrine[36]. While this SV was successfully captured by lrWGS data, srWGS failed to identify this amplification. No gene amplifications were observed for other drug resistance markers.

The results demonstrate enhanced sensitivity and resolution of SV calling pipelines leveraging lrWGS data, compared to approaches using srWGS data in isolation. Genome-wide SV maps for isolates BB12 and XHA, constructed using SVs concordant across Sniffles[33] (based on

lrWGS mapping to a reference genome) and Assemblytics[34] (based on *de novo* assemblies) are shown in Figures 6E and 6F respectively.

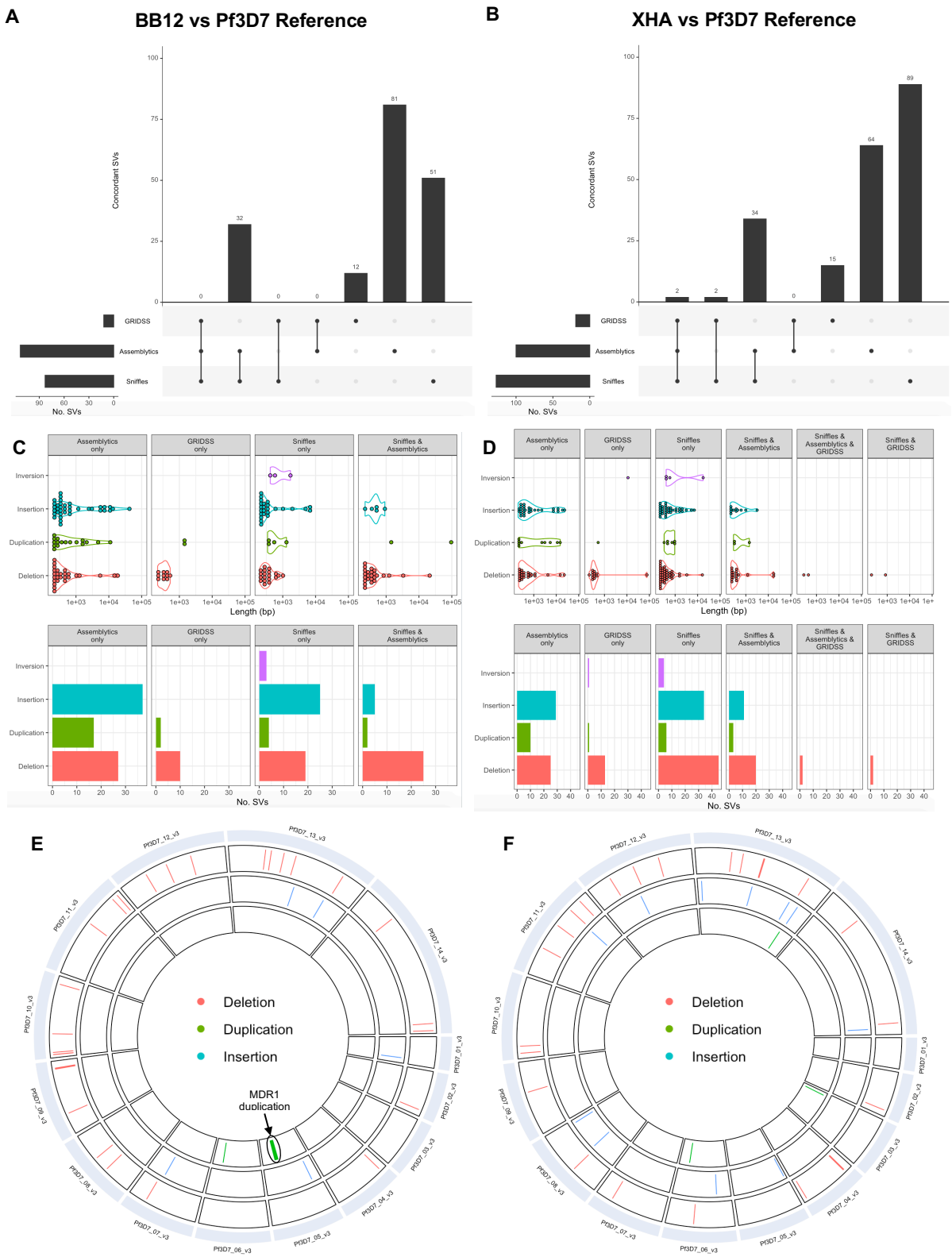


Figure 6: Comparison of structural variant calling piplines using long reads only (Sniffles), short reads only (GRIDSS) and our de novo assemblies (Assemblytics). *Overlaps in the sets of high-quality SVs identified by each pipeline are BB12 and XHA are shown in (A) and (B) respectively. SVs identified by different pipelines are considered equivalent if they overlap by at least 1bp, correspond to the same SV class and have similar lengths (i.e. the length of the shorter SV is at least 75% of that of*

the longer SV) (Text S8). Sniffles (which uses long reads only) and Assemblytics (which is informed by our de novo assemblies) show reasonably high concordance, while GRIDSS (which uses short reads only) exhibits very little concordance with either Sniffles or Assemblytics. SV types and lengths for BB12 and XHA are shown in (C) and (D) respectively. SV length and type distributions are generally quite similar for Sniffles and Assemblytics. Genome-wide SV maps, restricted to SVs concordant across Sniffles and Assemblytics, are shown for isolates BB12 (E) and XHA (F) respectively. Both Sniffles and Assembytics detected the known MDR1 duplication in BB12.

DISCUSSION

The Oxford Nanopore MinION sequencer is an attractive platform for sequencing malaria parasites in the field and laboratory. However it is known to have higher error rates in comparison to Illumina platforms. In addition, *P. falciparum* has an AT rich genome, with a higher probablility of homopolymers, known to be the cause of high error rates in Nanopore WGS. Here we benchmarked Nanopore long reads against Illumina short reads to reveal the utility of Nanopore for WGS of *P. falciparum* laboratory and field isolates. The results reveal a smaller baseline error rate than expected for haploid SNP genotyping using Nanopore long read sequencing. Discordant allele calls between Nanopore and Illumina data were more frequent in homopolymer regions greater than 6 bp, which are known to have poor sequence quality using the R9 (R9.4 and R9.5) Nanopore flow cells used here, and whose occurrence is expected to decrease with new Nanopore pore designs, as this is an active area of development. We optimised an enrichment and library preparation for whole blood-derived DNA samples enriched for parasite DNA after treatment with a *Mcr*BC digestion and rWGA, which could also be used for other *Plasmodium spp.* such as *P. vivax* and even other pathogens. Through sequencing field samples with a range of infection densities, we provide recommendations for selecting *P. falciparum* isolates for Nanopore WGS. While early runs resulted in outputs of 2-3 Gb, at the time of writing, we are generating up to 18 Gb of data per run using the more recently available SQK-LSK109 kit (unpublished data), which would increase coverage by almost ten fold and therefore allow multiplexing of up to 12 samples. In addition, we report streamlined analytical pipelines for SNP genotyping and de novo assembly, which yielded complete chromosomes including highly polymorphic subtelomeric regions and large SVs.

Removing dominant sources of variation (namely, indels; subtelomeric polymorphisms and heterozygosity) significantly increases the accuracy of genotyping with the Nanopore platform. The low error rates (<1% off filtered calls) computed were based on haploid SNP genotyping and are significantly lower than reported error rates for diploid or 'double haploid' SNP genotyping, when heterozygous calls are taken into account[29]. In addition, transition-type errors (G to A or C to T), were more frequent than transversion-type errors, possibly reflecting errors in basecalling. While the accuracy of Nanopore reference calls was high, alternate allele calls were found to be less reliable at low depths of coverage; however, overall discordance rates remained low since reference calls far outnumbered alternate calls. Reliably detecting alternate alleles at low depths of coverage may be problematic for Nanopore sequencing in some applications.

The results demonstrate the utility of the Nanopore MinION platform as a tool for drug-resistance profiling, even at very low depths of coverage (breadth of coverage ≥65% at 1X), granted caution is exercised around potential homopolymeric tracts flanking target loci where error rates were higher. High concordance with Illumina data was observed except for *crt* codons 72-76, which contains the chloroquine resistance haplotype. XHA (SVMNT) showed high concordance between Nanopore and Illumina data, but there were several discordant positions for BB12 (CVIET), due to a T to A mutation within codon 75 that creates a homopolymer. As these haplotypes are more common in some parts of the world, caution is

warranted when genotyping *crt* using Nanopore sequencing. Some drug resistance genes and samples were genotyped more consistently, e.g. *k13* was fully genotyped in almost all samples, whilst *crt* had more missing data due to the introduced homopolymer. With increasing information on the variant frequency in target populations however, missing genotypes can be imputed[37] and flow cell chemistry has improved to allow more accurate sequencing of homopolymers. Despite missing data for some samples, a large number of PNG isolates from different time points were successfully genotyped for all genes. We observed a high prevalence of resistance markers for chloroquine and antifolate drugs[38, 39] yet a complete lack of *kelch13* mutations associated with artemisinin resistance. This is critical information for PNG given the recent reports of a small number of C580Y mutant parasites in Wewak in 2016[40], only 50 km from the collection site for the samples sequenced here. The additional ability for Nanopore to return results within 48 hrs at field sites, rather than weeks or months after samples have often left the country of collection means that field relevant results can be returned in a time frame relevant for decision making regarding local treatment regimens or programmatic interventions.

As platform technologies change and improve, it will be important to ensure that new data can be directly compared to other, previously generated and public datasets[24], in order to track parasite evolution in response to changes in transmission and different selection pressures. While our analysis of discordance rates between Nanopore MinION and Illumina (haploid) SNP genotyping suggests the presence of some systematic differences between the two sequencing platforms, these effects may not be of a sufficient magnitude to significantly skew population genomics analyses. Our results are a preliminary validation, suggesting the potential viability of population genomics analyses combining Nanopore and Illumina data.

We also demonstrate the ability of Nanopore long read sequencing data to resolve hypervariable and low complexity genomic regions that have been difficult to characterise with short reads. Nanopore data allowed the construction of highly continuous genomic scaffolds that could then be polished with Illimina data to correct small-scale errors, including indels, individual base errors, block substitution events and local misassemblies. We generated highly complete and continuous *de novo* assemblies, validated through a comparison of BB12 against its ancestral reference strain IT4[19, 20, 41]. Although a number of *de novo* assemblies for *P. falciparum* drawing on single-molecule real-time (SMRT) long read sequencing using the PacBIO platform have been published[41], we presented the first *de novo* genome assemblies for *Plasmodium falciparum* generated using Nanopore sequence data.

While structural variants (SVs) play an important role in the genomic diversity of *P. falciparum*[42], shortcomings persist in the detection of SVs with short-read sequencing[32]. Long-read sequencing has the potential to enhance the sensitivity and specificity of SV detection, however, applications of long-read sequencing to characterise structural variation in the *P. falciparum* genome[43] have been limited. Studies have focused on comparisons between reference genomes and *de novo* genome assemblies generated from PacBIO SMRT long-read sequencing data. We demonstrated the enhanced ability of Nanopore long reads to capture SVs relative to short reads alone by comparing pipelines for SV detection relying on Illumina only mapped to a reference genome[35]; Nanopore only mapped to a reference genome[33], as well as *de novo* genome assemblies[34]. Of note, we detected a 96kbp duplication on chromosome 5 spanning *mdr1* in BB12, that was not detected by short read sequencing data. The ampliciation of *mdr1*, which is associated with multidrug resistance[36], was confirmed by qPCR. While further research is required to generate robust sensitivity and specificity estimates for SV detection, Nanopore sequencing is a promising tool for fine-scale mapping of complex SVs in the *Plasmodium falciparum* genome.

Nanopore long read sequencing using the MinION portable device has been proposed as a useful platform for real time portable sequencing solutions, with applications to malaria

surveillance. Here we have optimised Nanopore sequencing protocols and data analysis for high quality *P. falciparum* WGS. The results described offer insight into the quality and utility of Nanopore long read sequencing, and its inherent limitations. Following filtering of low confidence variation we observed acceptably low error rates and very high concordance of resulting genome wide (haploid) SNP genotypes for both pure *P. falciparum* DNA and field isolates contaminated with high amounts of human DNA. Long reads generated using this platform improve whole genome assembly and the detection of hypervariable regions and SVs. Nanopore genomes are directly comparable to publically available Illumina genomes and reveal novel insights of practical importance for malaria control programs including population structure and drug resistance. Finally, the Nanopore sequencing platforms also offers the potential for improving research equality in the many countries still battling malaria by enabling sequencing in country, giving research power to local malaria researchers and national malaria control programs to inform programmatic decisions.

METHODS AND MATERIALS

Parasite isolates

Pure *Plasmodium falciparum* genomic DNA was obtained from three culture-adapted isolates including the reference strain 3D7; BB12, a descendant of the IT strain[19, 20], and XHA, a culture-adapted isolate from PNG[18]. Annotated genome assemblies are available for 3D7[44] and IT[41]. *P. falciparum*-infected blood samples (n=33) were collected in cross-sectional surveys in Papua New Guinea, where malaria transmission and parasite genetic diversity is high[45] and in Cambodia where transmission is low[46] with high levels of multidrug resistance[47]. The research was approved by the PNG Institute of Medical Research Institutional Review Board No. 1116, Medical Research Advisory Council of PNG No. 1121, Cambodian National Committee on Health Research No. 265 NECHR and Walter and Eliza Hall Institute of Medical Research Human Research Ethics Committee Nos. 1303 and 1504.

DNA extraction, parasite DNA enrichment and library preparation

For cultured *P. falciparum* lines, DNA extraction was performed using DNeasy Blood & Tissue Kit (Qiagen, Germany) according to the manufacturer's instructions. Extracted DNA was purified and concentrated as described previously[48], with SPRIselect beads (Beckman Coulter, Australia) used in lieu of Ampure XP beads. Briefly, 0.45 volume re-suspended beads were added to 1 volume of sample and incubated at room temperature for 10 minutes. After incubation, the bead/DNA mixture was placed on a magnetic rack for 5 minutes. The supernatant was removed without disrupting the DNA-bound beads, and the beads were then washed twice with 200 µl of 80% ethanol. After incubating from an additional 10 minutes at room temperature, the bound DNA was eluted with 60 µl of 10 mM Tris buffer. The quantity and quality of extracted DNA were measured using a Qubit Fluorometer (Invitrogen Life Technologies, USA) and Nanodrop2000 (Thermofisher, USA) respectively. Size distributions were then analysed using Tapestation (Agilent Technologies, USA). Based on these results, 0.2pmol of DNA was processed with the 1D genomic sequencing kit SQK-LSK108 following the manufacturer's instructions (ONT, UK).

To optimise library preparation for field samples, mock infections were prepared to mimic 1% parasitemia by spiking 2 ng (0.083ng/ µL) of 3D7 parasite DNA with 73 ng (3.066/ µL) of commercial human DNA in a total volume of 24 µL. This is approximately 6.35 parasites/human genome and equivalent to a 1% parasitaemia. Both methylation-dependent digested (*Mcr*BC) and undigested mock infections were tested to verify the efficacy of the enrichment protocol (described in detail below). Library preparation was done using the 1D genomic sequencing kit following the manufacturer's instructions (Cat # SQK-LSK108, ONT, UK), however, we assessed various adjustments to the post-whole genome amplification (WGA, see below) and end prep purification. Following WGA, we trialled two methods of purification: 2.5V ethanol

precipitation and 1.8V SPRI beads. Following end prep, we further tested 0.45V and 1V volumes of bead clean-up. The optimal protocol was found to involve; 1.8V SPRI bead purification after WGA and 0.45V bead purification following end prep. All field samples were subject to this protocol. Three to four field samples were multiplexed for each run by indexing with the native barcoding kit (Cat # EXP NBD103, ONT, UK).

For field isolates, DNA was extracted from dried blood spots using the FavorPrep™ 96-Well Genomic DNA Extraction Kit (Favorgen, Taiwan). In order to reduce costs and minimise the amount of human DNA contamination in the sequencing output, we developed a novel potentially universal enrichment strategy that is dependent on low methylation of the *Plasmodium* genome (less than 1%)[49, 50], relative to the high levels of methylation in the human genome (60-90%)[51]. The protocol involves 2-steps: a restriction digest that targets methylated cytosines followed by whole genome amplification (WGA) using the Phi DNA polymerase primed with random hexamers. Briefly, 6 µL of DNA was subject to restriction digestion with *Mcr*BC (New England Biolabs, US) at 37°C for 1 hour, and halted by incubation at 80 °C for 20 minutes. Gel electrophoresis confirmed a smear for human DNA alone, indicating that human DNA had been digested, and maintenance of high molecular weight parasite DNA. Then, 2 µL of each field sample was subject to whole genome amplification using the illustra Genomiphi V2 DNA amplification kit (GE Healthcare Life Sciences, Australia) for 2 hours at 30°C followed by heat inactivation at 65°C for 10 minutes. Amplified DNA samples were then purified using 1.8V bead purification (Beckman Coulter, Australia) and further treated with T7 Endonuclease I to remove branch structures produced during whole genome amplification. All samples were amplified in duplicate-triplicate and pooled prior to library preparation.

To assess the efficacy of enrichment for mock and natural infections, we performed duplex qPCR, targeting the *P. falciparum 18S rRNA* genes[52] and the human *Plat1* gene[53]. We combined 0.2 µL (150 nM) of both forward and reverse primers; 0.45 µL (350 nM) of Taqman probes; 6.5µL of Taqman Fast Advanced Mastermix (Applied Biosystem, USA) and 4 µL of target DNA, comprising of undiluted original DNA or a 1:100 dilution of the amplified product. Details of the relevant primers and probes are presented in Table S1. To quantify parasite and human copy numbers, standard curves were generated by preparing 10 fold dilution series of plasmid DNA containing parasite *18SrRNA* gene[54] and from human genomic DNA (Cat No. G304, Promega, Australia). Thermal cycling was performed on a Light Cycler 480 II (Roche, Switzerland). The reaction volume was heated to 95° for 15 minutes, followed by 40 cycles of 95° for 15 seconds and 60° for 1 minute.

Nanopore sequencing and raw data analysis

Sequencing runs, each spanning 48 hours, were performed using Nanopore flowcells MIN106/MIN107 (R9.4 and R9.5) and MinKNOW software V1.7.10-1.11.5 (ONT, UK). Basecalling and demultiplexing were performed using the ONT Albacore (V1 for isolate XHA and V2.1 for others) to obtain reads in *fastq* format. Adapters were trimmed from basecalled reads using *Porechop V0.2.1*[55] with default parameters. Read length and quality summaries were subsequently generated with *NanoPlot V0.16.4*[56]. Trimmed reads were then aligned to the *P. falciparum* 3D7 (version 3)[44] reference genome as well as a concatenated P. *falciparum* 3D7, human HG19 and P. *vivax* P01 reference[57], using *Minimap2 V2.4*[58] with the *map-ont* preset. *Samtools V1.7*[59] utilities were used to sort and index and resultant alignments to obtain sorted bam files. Coverage statistics and the proportions of parasite and human DNA were determined using *samtools depth* and several in-house helper scripts. Locus-wise alignment summaries were generated using *samtools mpileup* V1.7[59], using a base quality threshold of 7. Alignment summaries were then used to call haploid variants using the *bcftools multiallelic-caller* V1.8[60], in both genotyping and discovery modes. Indels were removed using *vcftools V0.1.13*[61], and functional annotation of the resultant SNPs was performed using SnpEff V4.1l[62]. Quality filtration based on the depth of coverage at each locus and the proportion of reads

supporting the called allele was performed using an in-house script (Text S1). A schematic of this pipeline for processing Nanopore data is shown in Figure S1A.

Statistical analysis to determine associations between sample parameters and Nanopore sequencing output was done using the R package *ggpubr*[63]. Linear regression lines, with 95% confidence intervals, were generated with the function *ggscatter*, and Pearson's correlation coefficient was computed with the *stat_cor* utility.

Illumina sequencing and raw data analysis

We also performed Illumina sequencing of cultured *P. falciparum* isolates 3D7 (unknown origin), BB12 (Brazil) and XHA (PNG) which contain only *Plasmodium* spp. DNA, in addition to three Cambodian field isolates which are contaminated with human DNA. These samples were used to benchmark MinION sequencing against the widely used Illumina sequencing approach. Library preparation for field isolates required a preliminary parasite DNA enrichment step (see above). Sequencing was performed as per the Truseq Nano DNA sample preparation protocol (Illumina Inc., USA). Briefly, 200 ng of DNA was subject to shearing, end-repair, A-tailing and adapter ligation, followed by enrichment with 15 cycles of PCR using BIORAD T100 Thermal Cycler (Australia). The mean insert size was analysed using Tapestation (Agilent Technologies, USA) with D1000 Screen Tape (Cat No. 5067-5582). Libraries were then sequenced using the Nextseq 500 platform (Illumina Inc., USA) generating 75 bp paired-end reads with 6-base index read. Data analysis was done using an in-house pipeline (http://github.com/bahlolab/pf_variant_calling_pipeline) in accordance with GATK best practices[64] (Text S2).

Data analysis

Baseline error rates

To quantify false discovery rates for *de novo* SNP characterisation, variant calling was performed in discovery mode for both Nanopore and Illumina data to detect novel (haploid) SNPs in our cultured 3D7 isolate, relative to the 3D7 reference genome. Sequencing data for both Nanopore and Illumina comprised of mock infections. For Nanopore, we retained SNPs with minimum depth of coverage 10X and at least 75% of reads supporting the called allele. Illumina SNPs were filtered in accordance with GATK best practices (that is, QD>=2, MQ>=40, FS<=60, SOR<=3, MQRankSum>=-12.5 and ReadPosRankSum>=-8). To avoid detecting true variation between our cultured line and the isogenic reference strain that had arisen due to mitotic recombination *in vitro*[22], we considered only coding SNVs in essential genes that would have been unlikely to differ between the cultured and reference strains (Text S3).

SNV calling

Validation of SNV genotyping was done by first comparing *P. falciparum* isolates sequenced in-house using both Nanopore and Illumina platforms. Variants were called at 742,365 validated SNV loci obtained from an in-house reanalysis of the MalariaGEN Pf3k dataset (release 5), pooled with an additional set of 149 field isolates from Papua New Guinea[25]. Discordance rates between Nanopore and Illumina genotypes for cultured lines 3D7, BB12 and XHA, and the two Cambodian field isolates were quantified using an in-house script. Error rates in homopolymeric stretches more than 6bp in length, identified using GATK VariantAnnotator (V3.5) were compared against non-homopolymeric stretches. Filtration parameters for haploid Nanopore SNV genotyping (depth of coverage at least 2X, 75% of reads supporting the called allele) were subsequently deduced.

Drug resistance profiling was then performed for field isolates with at least 1X coverage in 65% of the genome. Genotypes were then screened for known resistance-associated variants in

multiple genes including *crt* (PF3D7_0709000), *mdr1* (PF3D7_0523000), *dhfr* (PF3D7_0417200), *dhps* (PF3D7_0810800) *and kelch13* (PF3D7_1344700).

To determine whether there were any platform-specific effects in the SNV genotypes called from WGS data, we then performed a principal coordinates analysis (PCoA) of in-house Nanopore field isolate data (PNG) and Illumina field isolate data obtained from the MalariaGEN Pf3k/PNG dataset (Asia Pacific including PNG and Cambodia) based on 51,421 high-quality polymorphic SNVs that were present in those countries (Text S4).

De novo genome assembly

Raw Nanopore reads for BB12 and XHA were first assembled into a scaffold using *Canu V.1.3*[65] using default parameters. Five iterations of consensus polishing were then performed with *Racon V.0.5.1*[66], which used the mapping of uncorrected reads to the scaffold, computed using *minmap2 V2.4*[58] to generate consensus sequences. Draft assemblies, constructed from Nanopore data only, were further polished with Illumina sequencing data. Illumina reads processed with *Trim Galore V.0.4.4*[67] were first mapped against draft assemblies using *bwa mem V.0.7.13*[68]. Individual base errors, indels, block substitution events, gaps, local misassemblies and ambiguous bases in the draft assemblies were then corrected using *Pilon V.1.22*[69] to obtain hybrid Nanopore-Illumina assemblies. A schematic of this pipeline is shown in Figure S1B. Draft assemblies for BB12 were benchmarked against its ancestral reference strain IT (Text S5).

Hybrid assemblies were compared against *P. falciparum* reference genomes using *nucmer* (*MUMmer, V.3.1*[31]), configured to identify one-to-one mappings between de novo contigs and reference chromosomes. Gene annotation was then performed with the *Companion* pipeline (July 2019 web server version)[70]. Domain classification of annotated *var* gene candidates was performed using the *VarDom 1.0 Server*[71] (Text S6).

Structural variant (SV) calling

To assess the utility of Nanopore lrWGS for the characterisation of SVs in *P. falciparum*, we performed a comparison of three distinct SV-calling pipelines for cultured lines BB12 and XHA: (i) Short read SV-calling with GRIDSS[35], which combines split-read, read-pair and assembly approaches, (ii) long read SV-calling with Sniffles[33] and NGLMR[33], which employs a split-read approach, and (iii) SV detection through direct comparison of our *de novo* assemblies and a reference genome using Assemblytics[34] (see Text S7 for details). The rationale for using GRIDSS as a benchmark for short read SV calling was two-fold: in addition to combining a number of common approaches to short read SV calling, GRIDSS was found to provide high precision across a range of SV types in a recent evaluation of over 60 short read SV detection algorithms[72]. Concordance rates across the three methodologies were then computed (Text S7). Systematic indel errors in Nanopore basecalling can lead to an overrepresentation of small deletions in low complexity and homopolymers, hence, we considered only SVs with length at least 200bp, in line with general recommendations for Sniffles[33]. Several spuriously large SVs, spanning almost entire chromosomes in some cases, were detected by both GRIDSS and Sniffles. Hence, only SVs with length below 300,000 bp were retained.

Acknowledgements
We are grateful to the volunteer communities and field teams and laboratory staff of PNG Institute of Medical Research and Institut Pasteur Cambodia for their involvement in sample collections. Thank you also to Celine Barnadas for facilitating access to Cambodian samples and Stuart Lee for assistance with genomic analyses.

Funding

This research was funded by grants from the Australian National Health and Medical Research Council (NHMRC) GNT1163420 and Department of Foreign Affairs and Trade. Samples from PNG were collected with funding from the NIH NIAID International Centres of Excellence in Malaria Research (ICEMR) South West Pacific (U19 AI089686) and a Bill & Melinda Gates Foundation TransEPI grant. LJR, IM and MB are supported by NHMRC Research Fellowships (GNT1161627, GNT1155075, GNT1102971). DLG is supported by the NIH NIAD International Centres of Excellence in Malaria Research (ICEMR) Asia Pacific (U19 AI129392-01) and the NHMRC Australian Centre of Research Excellence in Malaria Elimination Grant 1134989. The authors acknowledge the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support Scheme (IRIISS).

REFERENCES:

1. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome biology* **15**, 538 (2014).

2. Quick J*, et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232 (2016).

3. Miotto O*, et al.* Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nature genetics* **45**, 648-655 (2013).

4. Croucher NJ*, et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics* **45**, 656-663 (2013).

5. Russell CA*, et al.* Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine* **26 Suppl 4**, D31-34 (2008).

6. Organisation WH. World malaria report 2019. (ed^(eds) (2019).

7. Oyola SO*, et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *Malar J* **15**, 597 (2016).

8. Sundararaman SA*, et al.* Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nat Commun* **7**, 11078 (2016).

9. Dalmat R, Naughton B, Kwan-Gett TS, Slyker J, Stuckey EM. Use cases for genetic epidemiology in malaria elimination. *Malar J* **18**, 163 (2019).

10. Runtuwene LR*, et al.* Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum. *Sci Rep* **8**, 8286 (2018).

11. Imai K*, et al.* An innovative diagnostic technology for the codon mutation C580Y in kelch13 of Plasmodium falciparum with MinION nanopore sequencer. *Malar J* **17**, 217 (2018).

12. Nsanzabana C*, et al.* Molecular assays for antimalarial drug resistance surveillance: A target product profile. *PLoS One* **13**, e0204347 (2018).

13. Quick J*, et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* **16**, 114 (2015).

14. Pomerantz A*, et al.* Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**, (2018).

15. Hamilton WL*, et al.* Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Res* **45**, 1889-1901 (2017).

16. Tonkin-Hill GQ*, et al.* The Plasmodium falciparum transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen-encoding var genes. *PLoS Biol* **16**, e2004328 (2018).

17. Arnot D. Unstable malaria in Sudan: the influence of the dry season. Clone multiplicity of Plasmodium falciparum infections in individuals exposed to variable levels of disease transmission. *Trans R Soc Trop Med Hyg* **92**, 580-585 (1998).

18. Hill DL*, et al.* Merozoite Antigens of Plasmodium falciparum Elicit Strain-Transcending Opsonizing Immunity. *Infect Immun* **84**, 2175-2184 (2016).

19. Gaida A, Becker MM, Schmid CD, Bühlmann T, Louis EJ, Beck HP. Cloning of the repertoire of individual Plasmodium falciparum var genes using transformation associated recombination (TAR). *PLoS One* **6**, e17782 (2011).

20. Rogerson SJ, Reeder JC, al-Yaman F, Brown GV. Sulfated glycoconjugates as disrupters of Plasmodium falciparum erythrocyte rosettes. *Am J Trop Med Hyg* **51**, 198-203 (1994).

21. Oyola SO*, et al.* Efficient depletion of host DNA contamination in malaria clinical sequencing. *J Clin Microbiol* **51**, 745-751 (2013).

22. Bopp SE*, et al.* Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genet* **9**, e1003293 (2013).

23. Zhang M*, et al.* Uncovering the essential genes of the human malaria parasite. *Science* **360**, (2018).

24. Project MPfC. Genomic epidemiology of artemisinin resistant malaria. *Elife* **5**, (2016).

25. Tessema SK*, et al.* Antibodies to Intercellular Adhesion Molecule 1-Binding Plasmodium falciparum Erythrocyte Membrane Protein 1-DBLβ Are Biomarkers of Protective Immunity to Malaria in a Cohort of Young Children from Papua New Guinea. *Infect Immun* **86**, (2018).

26. Bowden R*, et al.* Sequencing of human genomes with nanopore technology. *Nat Commun* **10**, 1869 (2019).

27. Laver TW*, et al.* Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci Rep* **6**, 21746 (2016).

28. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* **19**, 90 (2018).

29. Malmberg MM, Spangenberg GC, Daetwyler HD, Cogan NOI. Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (Brassica napus L.). *Sci Rep* **9**, 8688 (2019).

30. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. *Hum Mol Genet* **27**, R234-R241 (2018).

31. Kurtz S*, et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

32. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Frontiers in genetics* **10**, 426 (2019).

33. Sedlazeck FJ*, et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018).

34. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021-3023 (2016).

35. Cameron DL*, et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res* **27**, 2050-2060 (2017).

36. Sidhu AB, Uhlemann AC, Valderramos SG, Valderramos JC, Krishna S, Fidock DA. Decreasing pfmdr1 copy number in plasmodium falciparum malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. *J Infect Dis* **194**, 528-535 (2006).

37. Otienoburu SD*, et al.* An online mapping database of molecular markers of drug resistance in Plasmodium falciparum: the ACT Partner Drug Molecular Surveyor. *Malar J* **18**, 12 (2019).

38. Fidock DA*, et al.* Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol Cell* **6**, 861-871 (2000).

39. Gregson A, Plowe CV. Mechanisms of resistance of malaria parasites to antifolates. *Pharmacol Rev* **57**, 117-145 (2005).

40. Miotto O*, et al.* Emergence of artemisinin-resistant Plasmodium falciparum with kelch13 C580Y mutations on the island of New Guinea. *bioRxiv*, 621813 (2019).

41. Otto TD*, et al.* Long read assemblies of geographically dispersed. *Wellcome Open Res* **3**, 52 (2018).

42. Ribacke U*, et al.* Genome wide gene amplifications and deletions in Plasmodium falciparum. *Mol Biochem Parasitol* **155**, 33-44 (2007).

43. Moser KA*, et al.* Strains used in whole organism Plasmodium falciparum vaccine trials differ in genome structure, sequence, and immunogenic potential. *Genome Med* **12**, 6 (2020).

44. Gardner MJ*, et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511 (2002).

45. Kattenberg JH*, et al.* The epidemiology of Plasmodium falciparum and Plasmodium vivax in East Sepik Province, Papua New Guinea, pre- and post-implementation of national malaria control efforts. *Malar J* **19**, 198 (2020).

46. Sluydts V*, et al.* Efficacy of topical mosquito repellent (picaridin) plus long-lasting insecticidal nets versus long-lasting insecticidal nets alone for control of malaria: a cluster randomised controlled trial. *Lancet Infect Dis* **16**, 1169-1177 (2016).

47. Hamilton WL*, et al.* Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. *Lancet Infect Dis* **19**, 943-951 (2019).

48. Blount BA, Driessen MR, Ellis T. GC Preps: Fast and Easy Extraction of Stable Yeast Genomic DNA. *Sci Rep* **6**, 26863 (2016).

49. Ponts N*, et al.* Genome-wide mapping of DNA methylation in the human malaria parasite Plasmodium falciparum. *Cell Host Microbe* **14**, 696-706 (2013).

50. Choi SW, Keyes MK, Horrocks P. LC/ESI-MS demonstrates the absence of 5-methyl-2'-deoxycytosine in Plasmodium falciparum genomic DNA. *Mol Biochem Parasitol* **150**, 350-352 (2006).

51. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).

52. Rosanas-Urgell A*, et al.* Comparison of diagnostic methods for the detection and quantification of the four sympatric Plasmodium species in field samples from Papua New Guinea. *Malar J* **9**, 361 (2010).

53. Pinheiro MM*, et al.* Plasmodium knowlesi genome sequences from clinical isolates reveal extensive genomic dimorphism. *PLoS One* **10**, e0121303 (2015).

54. Koepfli C*, et al.* Sustained Malaria Control Over an 8-Year Period in Papua New Guinea: The Challenge of Low-Density Asymptomatic Plasmodium Infections. *J Infect Dis* **216**, 1434-1443 (2017).

55. Wick R. Porechop. (ed^(eds). GitHub.

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

57. Auburn S*, et al.* A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res* **1**, 4 (2016).

58. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

59. Li H*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

60. Danecek P, Schiffels S, Durbin R. Multiallelic calling model in bcftools (-m). (ed^(eds) (2016).

61. Danecek P*, et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

62. Cingolani P*, et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).

63. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. (ed^(eds). CRAN.

64. Poplin R*, et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178 (2018).

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

66. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746 (2017).

67. Krueger F. TrimGalore. (ed^(eds). GitHub.

68. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*, (2013).

69. Walker BJ*, et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

70. Steinbiss S*, et al.* Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* **44**, W29-34 (2016).

71. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput Biol* **6**, (2010).

72. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**, 117 (2019).

SUPPORTING MATERIALS

Razook and Mehra *et al*. Real time, field-deployable whole genome sequencing of malaria parasites using nanopore technology

SUPPORTING METHODS

Text S1: Details of Nanopore SNV Quality Filtration

MinION genotype calls were assessed individually for each sample using two locus-wise quality metrics:

- The number of high-quality bases mapping to the locus $(\mathrm{CovDepth})$, and
- The proportion of high-quality reads supporting the called allele $(\mathrm{PropReads})$.

Both of these metrics were extracted from the DP4 tag, as described in the VCF file specification, which encodes the breakdown of high-quality bases mapping to the forward/reverse reference/non-reference alleles:

$$\mathrm{DP4} = \mathrm{REF}_{\mathrm{forward}}, \mathrm{REF}_{\mathrm{reverse}}, \mathrm{NREF}_{\mathrm{forward}}, \mathrm{NREF}_{\mathrm{reverse}},$$

where $(\mathrm{N})\mathrm{REF}_{\mathrm{forward}}$ and $(\mathrm{N})\mathrm{REF}_{\mathrm{reverse}}$ denote the number of high-quality bases mapping to the forward and reverse (non-)reference alleles respectively. The DP4 was parsed using string manipulation functions in GNU *awk,* and quality metrics were calculated using the following formulae:

$$\mathrm{CovDepth} = \mathrm{REF}_{\mathrm{forward}} + \mathrm{REF}_{\mathrm{reverse}} + \mathrm{NREF}_{\mathrm{forward}} + \mathrm{NREF}_{\mathrm{reverse}}$$

$$\mathrm{PropReads} = \begin{cases} \dfrac{\mathrm{REF}_{\mathrm{forward}} + \mathrm{REF}_{\mathrm{reverse}}}{\mathrm{REF}_{\mathrm{forward}} + \mathrm{REF}_{\mathrm{reverse}} + \mathrm{NREF}_{\mathrm{forward}} + \mathrm{NREF}_{\mathrm{reverse}}} & \text{if REF call} \\ \dfrac{\mathrm{NREF}_{\mathrm{forward}} + \mathrm{NREF}_{\mathrm{reverse}}}{\mathrm{REF}_{\mathrm{forward}} + \mathrm{REF}_{\mathrm{reverse}} + \mathrm{NREF}_{\mathrm{forward}} + \mathrm{NREF}_{\mathrm{reverse}}} & \text{if ALT call} \end{cases}$$

Since the DP4 tag considers only high-quality bases, $\mathrm{CovDepth}$ was lower than depth of coverage encoded in the DP tag for some loci.

Text S2: Pipeline for Processing Illumina Data

Isolates sequenced in-house on the Illumina platform were analysed in accordance with GATK best practices. Briefly, Illumina adapters were identified using *Picard Tools MarkIlluminaAdapters V2.17.3*[1] and reads were aligned against the P. *falciparum* 3D7 (version 3) reference genome using *bwa mem V0.7.13*[2]. Indels were realigned using the *GATK IndelRealigner V3.5.0*[3], and duplicate reads were identified with *PicardTools MarkDuplicates.* Base quality score recalibration was the performed using the *GATK BaseRecalibrator*. Haploid variants were called using the GATK *HaplotypeCaller* in both DISCOVERY mode (which outputted all detected variants) and GENOTYPE_GIVEN_ALLELES mode (which genotyped isolates at a specific set of loci).

Text S3: Essential Genes for Baseline Error Rate Analysis

Essentiality of genes for *in vitro* asexual blood stage development can be quantified using the piggyBac insertion mutagenesis score (MIS), which ranges from 0 for essential genes to 1 for dispensible genes[4]. To quantify false discovery rates for Nanopore and Illumina SNP calling, we restricted our attention to novel coding variants detected in the top quantile of essential genes (n=1356), with MIS in the range 0 to 0.142. Coding effects of each SNP were predicted using the R package *VariantAnnotation* (V1.32.0)[5].

Text S4: Principal Coordinates Analysis (PCoA) of Isolates from PNG and South-East Asia

Nanopore-sequenced field isolates were genotyped at 742,365 high-quality SNP loci obtained through our in-house reanalysis of the MalariaGEN Pf3k/PNG dataset. Only Nanopore genotypes with depth of coverage at least 2X and a minimum of 75% of reads supporting the called genotype were retained. Single-sample VCF files for Nanopore-sequenced field isolates were merged into a multisample VCF file using *vcftools V0.1.13*[6]. Multisample VCF files for both the MalariaGEN Pf3k/PNG datasets and our Nanopore-sequenced field isolates were then converted into *gds* format[7] and imported into R statistical software[8].

Isolates with genotype missingness rates exceeding 30% were removed, yielding 10 MinION-sequenced field isolates (all originating from Papua New Guinea) and 1040 Illumina-sequenced field isolates. Variant-level filtration was then performed to retain only SNP loci with missingness <10%, leaving 298,112 variants. SNPs monomorphic across the filtered panel of field isolates were removed to yield 51,421 polymorphic high-quality SNPs. These filtration steps were performed using *SeqArray V1.22.3*[9].

Pairwise distances between the resulting genotypes, defined to be the proportion of differing SNP sites between each pair of isolates, were computed using the *dist-gene* function in the R package *ape V5.2*[10]; missing sites were excluded in pairwise comparisons. Principal coordinates analysis (PCoA) was performed

on the resultant pairwise distance matrix with the base R function *cmdscale* (V3.6.2) with default parameters to examine clustering patterns stratified by geographic origin and sequencing platform.

Text S5: Benchmarking of BB12 *de novo* Assemblies Against the IT Reference Genome

Since the cultured line BB12 is a descendent of the IT strain of *Plasmodium falciparum*, we expect the core nuclear genomes of BB12 and IT to exhibit a high degree of similarity. Our preliminary benchmark for *de novo* assembly was thus a contig-level alignment of a potential BB12 assembly to the IT4 reference genome (version 4)[11]. We sought to minimise indel and mismatch rates for draft BB12 assemblies compared (Figure S1B) to the IT4 reference genome, generating quality summaries for various pipeline stages using *QUAST V.5.0.2*[12] (web server, May 2018 version).

The initial scaffold, generated using *Canu V.1.3*[13] from raw Nanopore reads, had high continuity, with all but one nuclear chromosome spanned by a single contig and two spurious contigs; however, mapping between the scaffold and the IT4 reference genome was poor (Table M1). Five iterations of consensus polishing with *Racon V0.5.1*[14] substantially improved alignment against the IT reference genome; however, the indel error rate relative to the IT4 reference genome remained elevated. To correct small-scale errors and local misassemblies, we further polished this draft assembly with Illumina data using *Pilon V.1.22*[15]. The resultant hybrid Nanopore-Illumina assembly exhibited substantially less indel error, with 218 indels and 66 mismatches per 100kb (Table M1).

Table M1: Benchmarking of draft assemblies for BB12 against IT reference genome

	Canu only	Canu + Racon (x5)	Canu + Racon (x5) + Pilon
No. contigs	17	17	17
Total length (bp)	21,817,649	22,819,515	23,097,724
Largest contig (bp)	3,142,754	3,312,166	3,350,589
N50 length (bp)	1,520,627	1,576,308	1,595,563
N75 length (bp)	1,298,640	1,358,305	1,375,631
GC content (%)	19.78	19.42	19.47
Genome fraction (%)	33.564	96.619	96.764
Total aligned length	7,539,785	22,142,137	22,497,519
Mismatches (per 100kbp)	89.10	150.97	65.84
Indels (per 100 kbp)	2216.89	1269.48	217.96

Text S6: Gene Annotation and *var* Gene Domain Classification

Gene annotation of hybrid assemblies was performed with the *Companion* pipeline[16] via the web server (https://companion.sanger.ac.uk) (July 2019 version), using *P. falciparum 3D7* as a reference strain and taking into account reference protein evidence. Highly conserved genes were mapped using the Rapid Annotation Transfer Tool (RATT), while a relatively lenient threshold inclusion score of 0.5 was used for de novo gene prediction by AUGUSTUS, to increase the sensitivity of gene annotation. Pseudogene detection was also performed as part of the Companion pipeline.

To characterise the ability of Nanopore long reads to access complex genomic regions, genes encoding either PfEMP1, DBL-domains, N-terminal or acidic-terminal segments (all associated with *var* genes) were identified. Protein sequences of all potential full and partial *var* genes were extracted and classified using the *VarDom 1.0 Server*[17] (http://www.cbs.dtu.dk/services/VarDom/), with the default lower-bit score threshold 9.97 for homology blocks. Annotated *var* genes were considered to be complete if they included both acidic- and N-terminal segments. Accounting for the possible disruption of ORFs due to local assembly errors, neighbouring annotated *var* candidates within 1200bp windows were concatenated if their domain structures suggested they corresponded to the same gene to obtain 'split' *var* genes. *Var* candidates that included at least 3 annotated domains but did not contain either acidic- or N-terminal segments were also considered to be partial *var* genes.

Text S7: Structural Variant Calling Pipelines

For short read data (Illumina), adapter trimming was first performed with *TrimGalore*. Trimmed Illumina reads were aligned against the 3D7 reference genome with *bwa mem V0.7.13*[2], and resultant alignments files were sorted and indexed with *Samtools V1.7*[18] utilities. Structural variant calling was then performed with *GRIDSS V2.8.3*[19]. Unpaired breakend variants were removed, and only simple variant types (i.e. insertions, deletions, duplications and inversions) were annotated using the R package *StructuralVariantAnnotation V1.0.0*.

Structural variants were retained if they passed default quality filters (i.e. the FILTER flag in the VCF files generated by GRIDSS was set to PASS).

For long read data (Nanopore), reads were first aligned against the 3D7 reference genome using *NGMLR V0.2.7*[20] with the *ont* preset. The resultant alignments were sorted and indexed using *Samtools V1.7*[18] utilities. Structural variant calling was subsequently performed with *Sniffles V1.0.11*[20] using default parameters. Breakends were removed to retain only insertions, deletions, translocations, duplications and inversions. A threshold of 20 high-quality reads supporting the called variant (as determined by the RE tag in the output VCF file) was implemented to obtain high-quality structural variants. Overlapping duplications and deletions were assumed to be spurious and were accordingly removed.

To perform structural variant calling with our *de novo* assemblies, we first aligned our assemblies against 3D7 reference genome using *nucmer* (*MUMmer, V.3.1*[21]). All anchor matches were used, irrespective of their uniqueness, and a minimum cluster length of 500 was implemented; maximal exact matches were also required to be at least 100bp in length. The resultant delta files were processed using the Assemblytics web server[22] to identify structural variants.

To quantify concordance rates across these methodologies we compared the set of high-quality SVs detected by each pipline using a custom *R* script. SVs identified by different pipelines were considered to be equivalent if: (i) they overlapped by at least 1bp; (ii) were annotated within the same SV class (Table M2); and (iii) if the length of the shorter SV was at least 75% of the length of the longer SV. To account for systematic indel errors in MinION basecalling, particularly in homopolymeric tracts, we retained only SVs with length at least 200 bp[20]. SVs exceeding 300 kbp seemed to be spurious across all three pipelines and were thus removed.

Table M2: SV classes identified by each pipeline

SV Class	*Sniffles*	*GRIDSS*	*Assemblytics*
Deletion	DEL	DEL	Deletion, Repeat_contraction, Tandem_contraction
Insertion	INS	INS	Insertion, Repeat_expansion
Duplication	DUP, INVDUP	DUP	Tandem_expansion
Inversion	INV, INVDUP	INV	
Translocation	TRA	CTX	

REFERENCES

1. Institute B. Picard Tools. (ed^(eds). GitHub.

2. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*, (2013).

3. McKenna A*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

4. Zhang M*, et al.* Uncovering the essential genes of the human malaria parasite. *Science* **360**, (2018).

5. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076-2078 (2014).

6. Danecek P*, et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

7. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).

8. Team RC. R: A language and environment for statistical computing. (ed^(eds). R Foundation for Statistical Computing (2013).

9. Zheng X*, et al.* SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251-2257 (2017).

10. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019).

11. Otto TD*, et al.* Long read assemblies of geographically dispersed. *Wellcome Open Res* **3**, 52 (2018).

12. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).

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

14. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746 (2017).

15. Walker BJ*, et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

16. Steinbiss S*, et al.* Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* **44**, W29-34 (2016).

17. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput Biol* **6**, (2010).

18. Li H*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

19. Cameron DL*, et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res* **27**, 2050-2060 (2017).

20. Sedlazeck FJ*, et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018).

21. Kurtz S*, et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

22. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021-3023 (2016).

SUPPORTING FIGURES

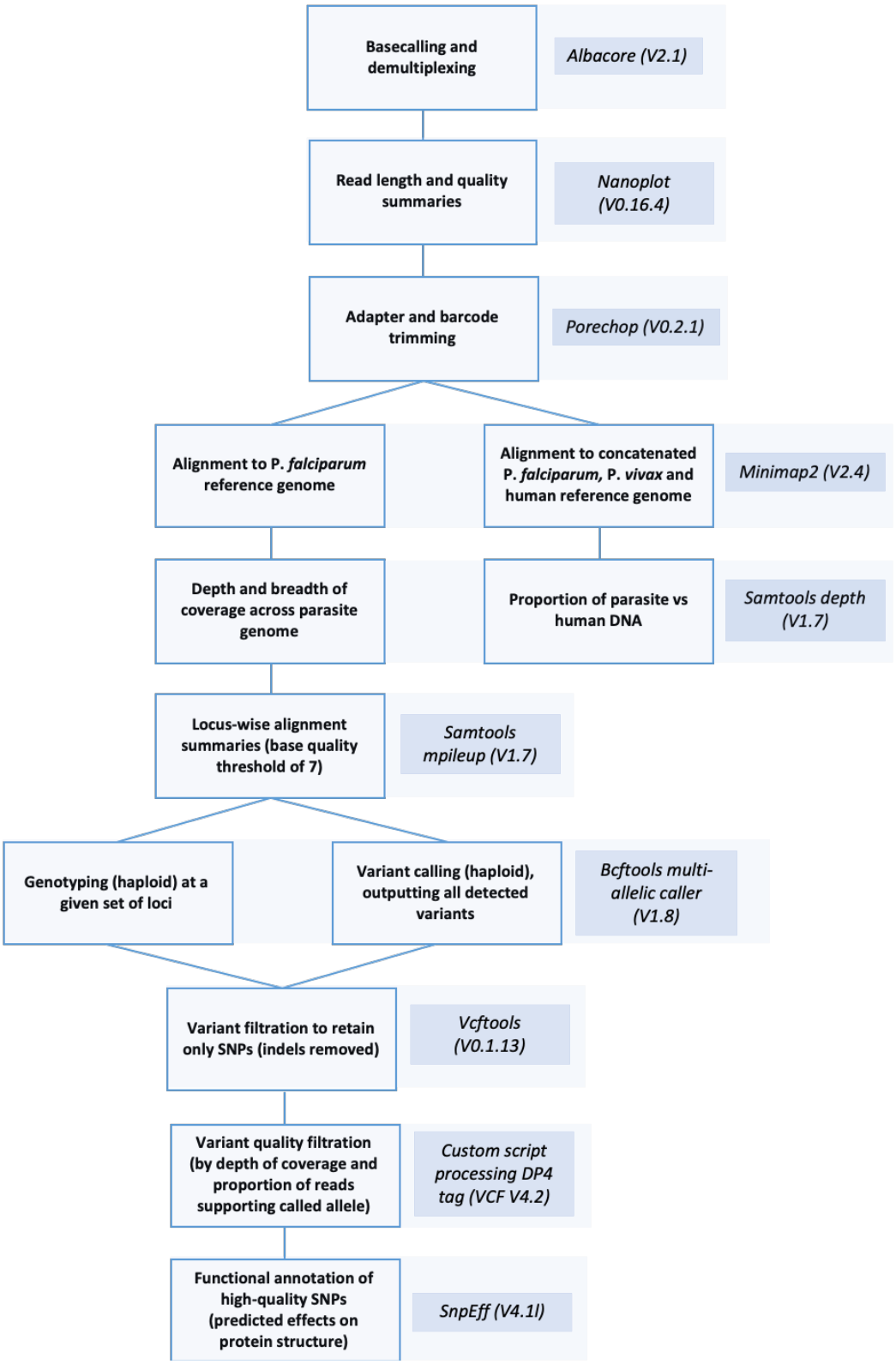


Figure S1A. Variant calling and alignment pipeline for Nanopore data

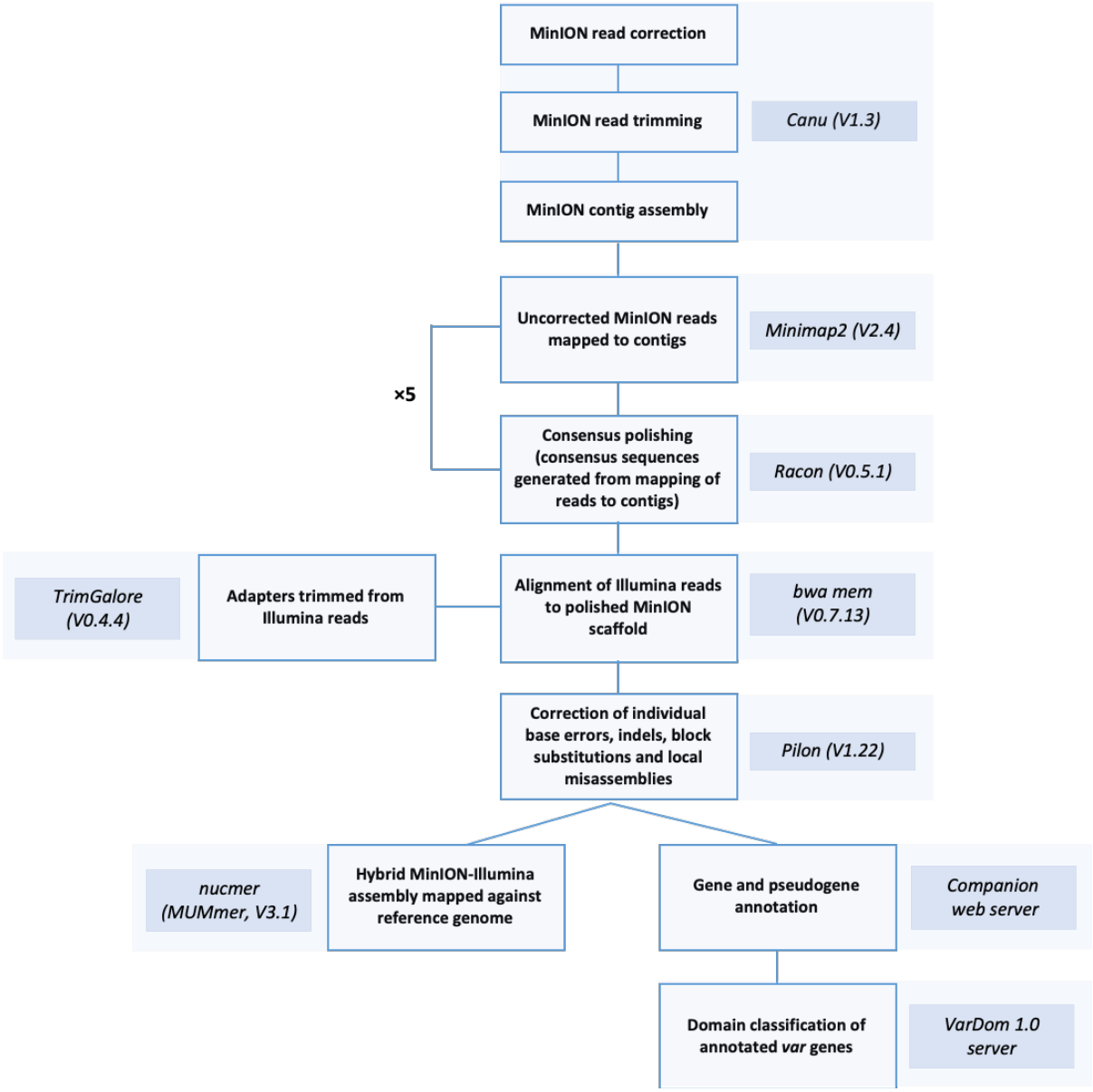


Figure S1B. *De novo* assembly pipeline for hybrid assemblies of Nanopore and Illumina data

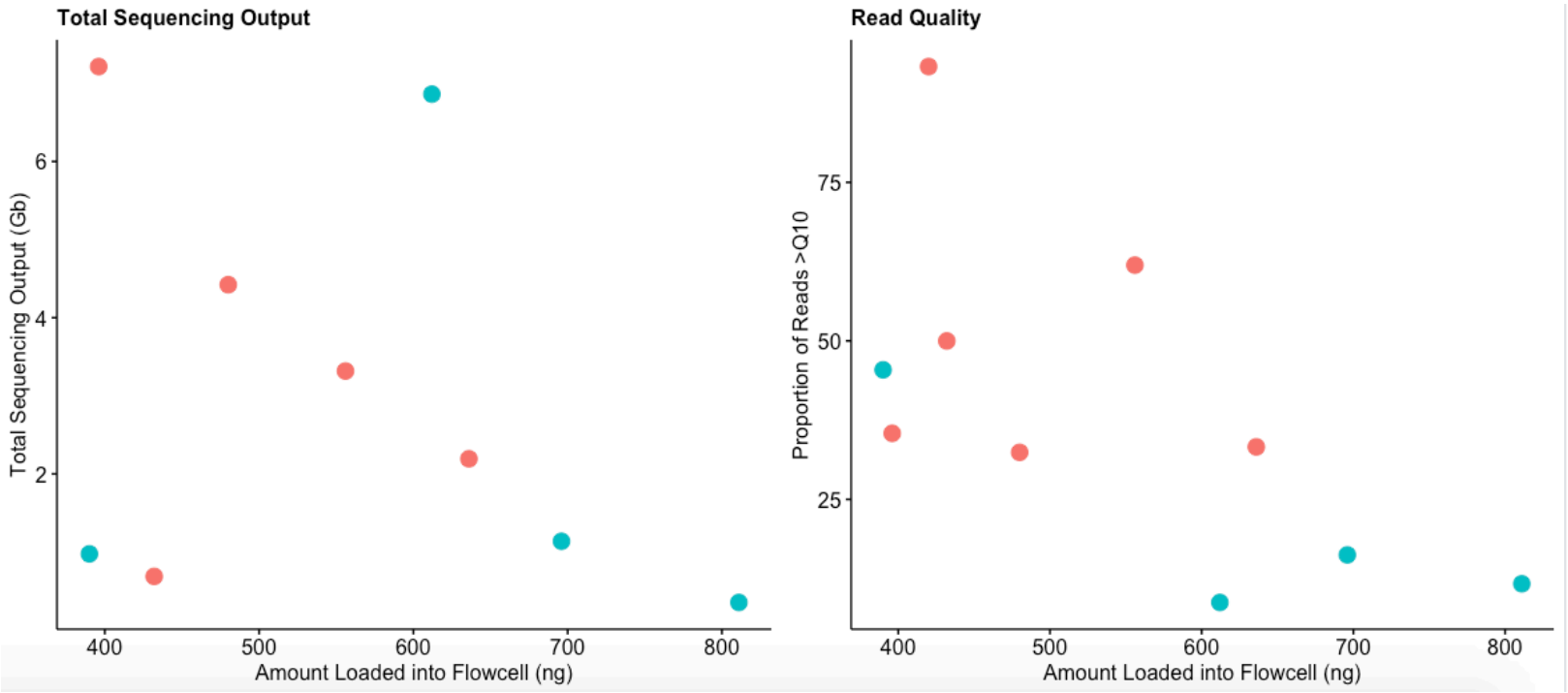


Figure S2. Parameters associated with Nanopore sequencing output, by flowcell. Nanopores tend to become saturated when large amounts of DNA are loaded into a flowcell, leading to a generally decreasing trend between the total sequencing output/quality and the amount of loaded DNA.

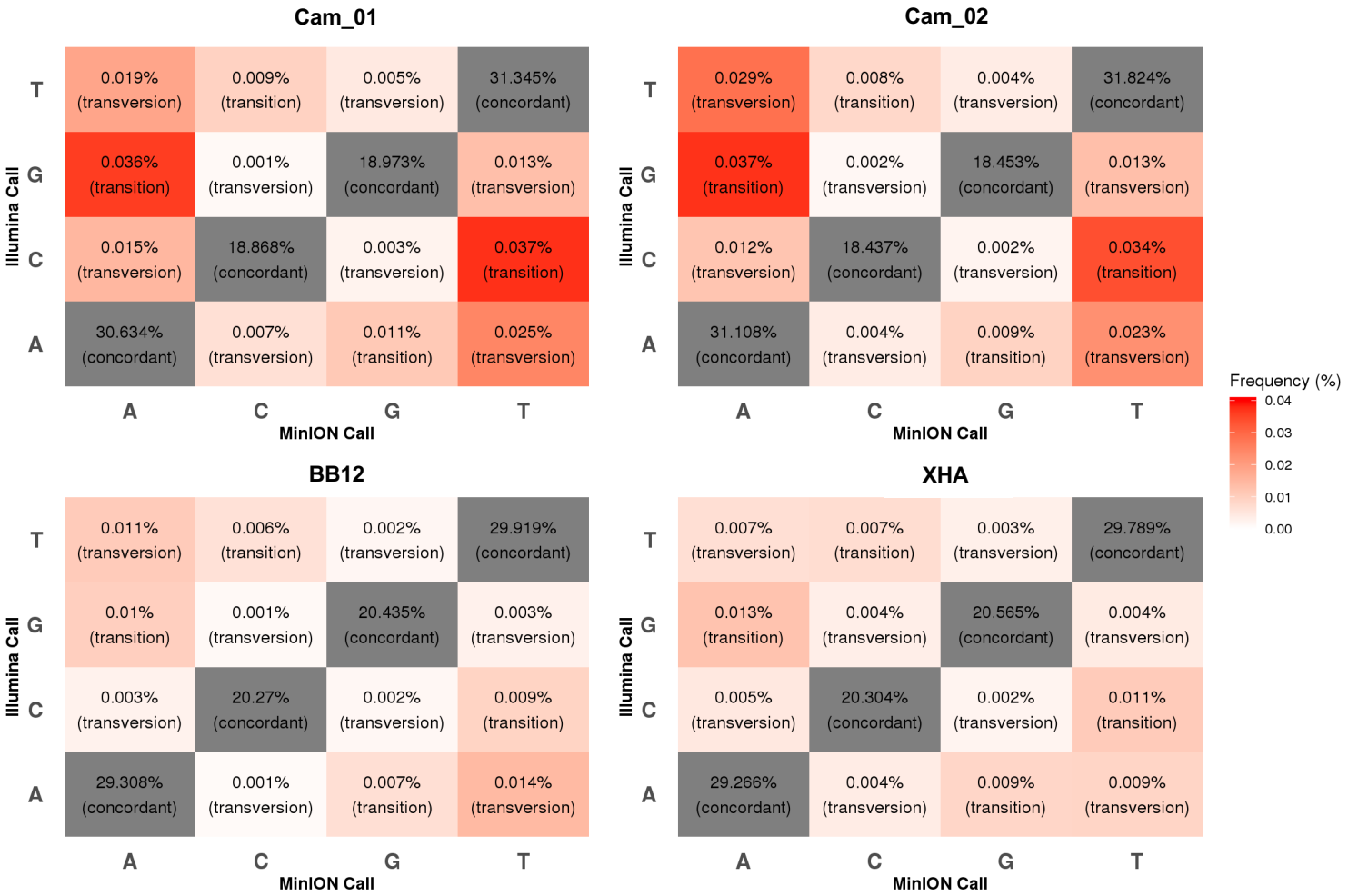


Figure S3. Relative frequencies of concordant and discordant genotype calls using Nanopore and Illumina data. Discordant calls were stratified by transition/transversion. Transition errors (between bases with similar ring structures) tend to be more common than transversion errors.

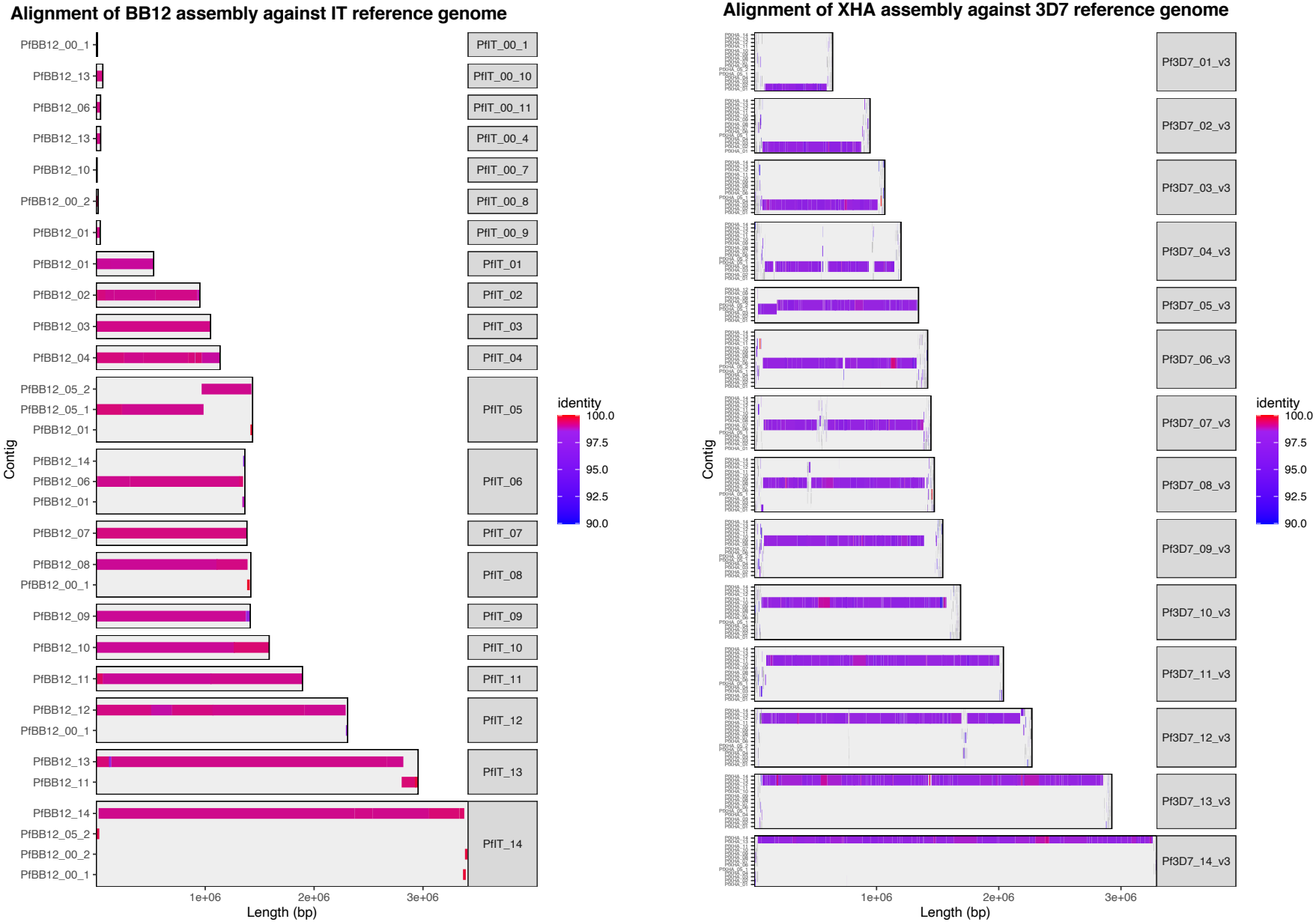


Figure S4. Alignments of *de novo* BB12 and XHA assemblies against reference genomes IT4 (version 4) and 3D7 (version 3) respectively, with one-to-one mappings between de novo contigs and reference chromosomes computed using nucmer (MUMmer, V.3.1). Assembly continuity is high, with all but one nuclear chromosome spanned by a single contig. De novo contigs and reference genomes are generally concordant in core genomic regions, but alignments tend to become fragmented in subtelomeric hypervariable region

A.

COMPLETE (n=27)

Contig	Dir	Start	End	Gene(s)	Domain Structure									
tg00000031	+	387210	398129	Pf_050041100	NTS	DBLa	CIDRa	DBLb	DBLg	DBLe	DBLg	DBLz	DBLe	ATS
tg00000027	-	1014328	1024713	Pf_030030800	NTS	DBLa	CIDRa	DBLd	CIDRg	DBLe	DBLz	DBLe	ATS	
tg00000020	-	38859	49502	Pf_070006100	NTS	DBLa	CIDRg	DBLb	DBLd	CIDRg	DBLz	DBLe	ATS	
tg00000004	-	48848	58897	Pf_120006700	NTSpam	DBLpam1	DBLpam2	CIDRpam	DBLpam3	DBLe	DBLe	DBLe	ATS	
tg00000039	-	3319224	3331125	Pf_140087500	NTS	DBLa	CIDRa	DBLb	DBLd	CIDRg	DBLg	DBLz	ATS	
tg00000029	+	19757	29476	Pf_020008600	NTS	DBLa	CIDRa	DBLb	DBLd	CIDRg	DBLg	ATS		
tg00000004	+	2254349	2263040	Pf_120062500	NTS	DBLa	CIDRb	DBLg	DBLz	DBLe	DBLe	ATS		
tg00000019	-	702626	711217	Pf_060022400	NTS	DBLa	CIDRa	DBLb	DBLg	DBLd	CIDRg	ATS		
tg00000023	+	29825	38421	Pf_050006200	NTS	DBLa	CIDRa	DBLb	DBLd	CIDRb	ATS			
tg00000004	+	1723341	1731362	Pf_120047200	NTS	DBLa	CIDRa	DBLe	DBLz	DBLe	ATS			
tg00000006	-	1976855	1985905	Pf_110054400	NTS	DBLa	CIDRa	DBLb	DBLd	CIDRb	ATS			
tg00000002	-	31911	41304	Pf_130006000	NTS	DBLa	DBLg	DBLe	DBLd	CIDRb	ATS			
tg00000027	+	20928	28104	Pf_030005200	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000016	+	397087	404890	Pf_080013800	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000006	+	22738	30018	Pf_110005700	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000004	+	26055	34261	Pf_120006200	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000004	+	1695282	1702784	Pf_120046700	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000004	+	1714428	1721923	Pf_120047100	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000025	-	873055	879737	Pf_040028600	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000025	-	912273	920409	Pf_040029100	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000025	-	1106037	1113680	Pf_040034300	NTS	DBLa	CIDRa	DBLd	CIDRg	ATS				
tg00000031	-	419356	426549	Pf_050041900	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000004	-	776658	784284	Pf_120025500	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000002	-	2856518	2863633	Pf_130078700	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000033	-	560062	567221	Pf_010019500	NTS	DBLa	CIDRa	DBLb	ATS					
tg00000025	-	884420	889761	Pf_040028700	NTS	DBLa	CIDRa	ATS						
tg00000039	-	61109	65241	Pf_140006500	NTS	DBLa	DBLe	ATS						

SPLIT (n=16)

Contig	Dir	Start	End	Gene(s)	Domain Structure									
tg00000027	+	1002348	1013225	Pf_030030600,Pf_030030700	NTS	DBLa	CIDRa	DBLb	DBLg	DBLg	DBLd	CIDRb	791bp	ATS
tg00000019	+	1345115	1356162	Pf_060038200,Pf_060038300	NTS	DBLa	CIDRa	DBLb	DBLb	DBLg	DBLd	CIDRb	704bp	ATS
tg00000042	+	1539652	1550184	Pf_100044600,Pf_100044700	NTS	DBLa	CIDRa	DBLb	DBLg	DBLg	DBLd	CIDRb	675bp	ATS
tg00000020	-	506384	517153	Pf_070017300,Pf_070017200	NTS	DBLa	CIDRa	DBLb	DBLd	CIDRg	DBLg	DBLz	651bp	ATS
tg00000042	-	1551316	1561180	Pf_100044900,Pf_100044800	NTS	DBLa	CIDRa	DBLb	DBLg	DBLd	CIDRb	981bp	ATS	
tg00000020	-	548018	557889	Pf_070018000,Pf_070017900	NTS	DBLa	CIDRa	25bp	DBLg	DBLd	CIDRb	ATS		
tg00000004	-	765441	775082	Pf_120025400,Pf_120025300	NTS	DBLa	CIDRa	DBLb	83bp	DBLd	CIDRb	ATS		
tg00000020	+	28789	37802	Pf_070005900,Pf_070006000	NTS	145bp	CIDRa	DBLb	DBLd	CIDRb	ATS			
tg00000004	+	1757379	1764810	Pf_120048000,Pf_120048100	NTS	42bp	DBLa	CIDRa	DBLd	CIDRb	ATS			
tg00000025	-	533267	540934	Pf_040020600,Pf_040020500	NTS	42bp	DBLa	CIDRa	DBLd	CIDRb	ATS			
tg00000025	-	896785	904808	Pf_040028900,Pf_040028800	NTS	DBLa	CIDRa	DBLd	CIDRb	212bp	ATS			
tg00000019	-	1357247	1366157	Pf_060038600,Pf_060038400	NTS	885bp	CIDRa	DBLg	DBLd	CIDRb	ATS			
tg00000020	-	537548	546613	Pf_070017800,Pf_070017700	NTS	DBLa	CIDRa	DBLb	56bp	CIDRb	ATS			
tg00000004	-	2264049	2271729	Pf_120062700,Pf_120062600	NTS	DBLa	CIDRa	DBLd	CIDRb	101bp	ATS			
tg00000016	+	412725	420025	Pf_080014000,Pf_080014100	NTS	216bp	CIDRa	DBLd	CIDRb	ATS				
tg00000004	+	1737785	1745602	Pf_120047400,Pf_120047500,Pf_120047700	NTS	DBLa	94bp	DBLd	480bp	ATS				

PARTIAL (n=11)

Contig	Dir	Start	End	Gene(s)	Domain Structure								
tg00000002	+	20623	30908	Pf_130005900	–	DBLa	CIDRa	DBLb	DBLd	CIDRg	DBLg	DBLz	ATS
tg00000020	-	519887	528754	Pf_070017500,Pf_070017400	–	CIDRa	DBLb	DBLg	DBLd	CIDRb	952bp		
tg00000020	+	559875	568672	Pf_070018100	–	DBLa	CIDRd	DBLb	DBLg	DBLg	ATS		
tg00000025	+	29805	34781	Pf_040008200	NTS	DBLa	CIDRa	DBLd	CIDRb	–			
tg00000039	+	39615	44859	Pf_140006200	NTS	DBLa	CIDRa	DBLd	CIDRg	–			
tg00000033	-	573756	576278	Pf_010020100,Pf_010020000,Pf_010019900	NTS	26bp	DBLa	125bp	CIDRa	–			
tg00000025	-	508033	513473	Pf_040020200	NTS	DBLa	CIDRa	DBLd	CIDRg	–			
tg00000025	-	523080	528397	Pf_040020400	NTS	DBLa	CIDRa	DBLd	CIDRb	–			
tg00000020	-	572106	577367	Pf_070018200	NTS	DBLa	CIDRa	DBLd	CIDRb	–			
tg00000016	-	1356800	1361793	Pf_080038100	NTS	DBLa	CIDRa	DBLd	CIDRb	–			
tg00000029	-	904465	907349	Pf_020031500	NTS	DBLa	CIDRa	–					

B.

COMPLETE (n=18)

Contig	Dir	Start	End	Gene(s)	Domain Structure									
tg00000013	-	546257	557851	PfXHA_070018300	NTS	DBLa	CIDRa	DBLb	DBLg	DBLd	CIDRb	DBLb	DBLg	ATS
tg00000021	+	1030844	1041909	PfXHA_030031500	NTS	DBLa	CIDRa	DBLb	DBLb	DBLg	DBLd	CIDRb	ATS	
tg00000013	+	37217	48841	PfXHA_070006100	NTS	DBLa	CIDRa	DBLb	DBLd	DBLe	DBLz	DBLe	ATS	
tg00000011	+	1558431	1568495	PfXHA_100045800	NTS	DBLa	CIDRg	DBLg	DBLe	DBLe	DBLz	DBLe	ATS	
tg00000018	+	30296	39580	PfXHA_040006000	NTS	DBLa	CIDRa	DBLb	DBLg	DBLe	DBLe	ATS		
tg00000022	-	40802	49789	PfXHA_020006500	NTS	DBLa	CIDRg	DBLg	DBLe	DBLz	DBLe	ATS		
tg00000013	-	535465	544837	PfXHA_070018200	NTS	DBLa	CIDRa	DBLb	DBLg	DBLz	DBLe	ATS		
tg00000006	-	31586	40888	PfXHA_110006200	NTS	DBLa	DBLb	DBLg	DBLz	DBLe	ATS			
tg00000022	+	32205	39752	PfXHA_020006400	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000026	+	30809	39061	PfXHA_050005800	NTS	DBLa	CIDRa	DBLz	DBLe	ATS				
tg00000004	+	20145	27651	PfXHA_120005600	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000004	+	1702252	1709913	PfXHA_120047000	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000018	-	529655	537112	PfXHA_040018500	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000018	-	543735	550646	PfXHA_040018800	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000018	-	887073	893998	PfXHA_040026900	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000018	-	900996	908480	PfXHA_040027000	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000011	-	1569528	1576839	PfXHA_100045900	NTS	DBLa	CIDRa	DBLd	CIDRb	ATS				
tg00000007	-	40449	44425	PfXHA_090005900	NTS	DBLa	DBLe	ATS						

SPLIT (n=16)

Contig	Dir	Start	End	Gene(s)	Domain Structure									
tg00000006	-	2013307	2023745	PfXHA_110056400,PfXHA_110056300	NTS	DBLa	CIDRa	DBLd	CIDRg	DBLz	DBLe	47bp	DBLe	ATS
tg00000003	+	2865002	2875282	PfXHA_130082400,PfXHA_130082500,PfXHA_13008260	NTSpam	DBLpam1	DBLpam2	86bp	DBLpam3	DBLe	41bp	DBLe	ATS	
tg00000004	-	755457	765304	PfXHA_120025600,PfXHA_120025500,PfXHA_12002540	NTS	DBLa	CIDRa	78bp	CIDRg	DBLz	DBLe	728bp	ATS	
tg00000009	+	438011	445181	PfXHA_080014400,PfXHA_080014500,PfXHA_08001460	NTS	DBLa	CIDRa	53bp	DBLd	CIDRb	974bp	ATS		
tg00000018	-	509413	518281	PfXHA_040018200,PfXHA_040018100	NTS	DBLa	CIDRa	DBLd	48bp	CIDRg	DBLg	ATS		
tg00000018	-	933994	941397	PfXHA_040027700,PfXHA_040027600,PfXHA_04002750	NTS	DBLa	47bp	CIDRa	104bp	DBLd	CIDRb	ATS		
tg00000013	-	49883	57659	PfXHA_070006400,PfXHA_070006300,PfXHA_07000620	NTS	DBLa	CIDRd	119bp	DBLz	82bp	DBLe	ATS		
tg00000014	+	31216	38915	PfXHA_060006000,PfXHA_060006100,PfXHA_06000620	NTS	DBLa	CIDRa	113bp	CIDRb	35bp	ATS			
tg00000009	+	446609	453754	PfXHA_080014700,PfXHA_080014800	NTS	DBLa	CIDRa	DBLb	DBLg	27bp	ATS			
tg00000004	+	1760468	1768339	PfXHA_120048000,PfXHA_120048100	NTS	DBLa	CIDRa	DBLd	CIDRb	965bp	ATS			
tg00000022	-	894075	901932	PfXHA_020028800,PfXHA_020028700	NTS	DBLa	CIDRa	41bp	DBLd	CIDRb	ATS			
tg00000018	-	909903	917526	PfXHA_040027200,PfXHA_040027100	NTS	DBLa	CIDRa	DBLd	CIDRb	1099bp	ATS			
tg00000009	+	461595	469409	PfXHA_080015000,PfXHA_080015100	NTS	DBLa	CIDRa	38bp	CIDRb	ATS				
tg00000024	-	624400	632131	PfXHA_010022700,PfXHA_010022500	NTS	DBLa	823bp	DBLd	CIDRb	ATS				
tg00000013	-	503845	511830	PfXHA_070017500,PfXHA_070017400	NTS	DBLa	78bp	DBLd	CIDRb	ATS				
tg00000013	-	519595	526993	PfXHA_070017900,PfXHA_070017700	NTS	DBLa	CIDRa	DBLd	794bp	ATS				

PARTIAL (n=15)

Contig	Dir	Start	End	Gene(s)	Domain Structure									
tg00000004	-	35603	45661	PfXHA_120006000,PfXHA_120005900	–	DBLpam1	DBLpam2	CIDRpam	DBLpam3	DBLe	DBLe	DBLe	449bp	ATS
tg00000004	+	2227867	2235876	PfXHA_120061600	NTSpam	DBLpam1	DBLpam2	CIDRpam	DBLpam3	DBLe	DBLe	DBLe	–	
tg00000014	-	40698	51096	PfXHA_060006400,PfXHA_060006300	NTS	DBLa	CIDRa	DBLb	DBLg	121bp	DBLd	CIDRb	–	
tg00000009	-	1428724	1436962	PfXHA_080040100	NTS	DBLa	CIDRa	DBLb	DBLg	DBLd	CIDRb	DBLb	–	
tg00000003	-	27257	35470	PfXHA_130005700	NTS	DBLa	CIDRa	DBLb	DBLg	DBLd	CIDRb	DBLb		
tg00000003	+	37206	45155	PfXHA_130005800,PfXHA_130005900	–	DBLb	DBLb	DBLd	CIDRb	923bp	ATS	–		
tg00000024	+	45147	50224	PfXHA_010006700	NTS	DBLa	CIDRa	DBLd	CIDRb	–				
tg00000004	+	1683153	1688338	PfXHA_120046600	NTS	DBLa	CIDRa	DBLd	CIDRg	–				
tg00000018	-	927694	932618	PfXHA_040027400	NTS	DBLa	CIDRa	DBLd	CIDRb	–				
tg00000003	-	2882350	2887594	PfXHA_130082800	NTS	DBLa	CIDRa	DBLd	CIDRb	–				
tg00000006	+	25165	30487	PfXHA_110006100	–	DBLd	CIDRb	ATS						
tg00000001	+	29641	34407	PfXHA_140005900	–	DBLd	CIDRb	ATS						
tg00000014	-	745910	748894	PfXHA_060024500	NTS	DBLa	CIDRa	–						
tg00000013	-	564692	566983	PfXHA_070018700	NTS	DBLa	CIDRa	–						
tg00000004	-	2250204	2256010	PfXHA_120062100	–	DBLd	CIDRg	ATS						

Figure S5. Domain structures for annotated *var* genes. *Var* gene structures were identified after *de novo* assembly of reads for *P. falciparum* laboratory isolates A) BB12 B) XHA

SUPPORTING TABLES

Table S1. Primers and probes used in duplex qPCR assay
Table S2. Results of sequencing experiments
Table S3. SNV genotyping concordance rates for two laboratory and two field isolates in homopolymeric regions

Table S1. Primers and Probes used in Duplex qPCR Assay

Species	Primer/ Probe	Sequence detail (5' to 3')	Reference
P.falciparum	Fal_F	TATTGCTTTTGAGAGGTTTTGTTACTTTG	Rosanas-Urgell et al., 2010
	Fal_R	ACCTCTGACATCTGAATACGAATGC	
	Fal_P	FAM- ACGGGTAGTCATGATTGAGTT - MGB-NFQ	
Human	Hu_F	CTTACCACATCCGCTCCATC	Pinheiro, M.M., et al., 2015
	HU_R	TTCACACTCTCCGTCACATTG	
	HU_P	LC640 -CACATCCCCAGTGCCGAGTTAGA- BBQ	

Table S2A. Nanopore Sequencing Output for Pure Cultured Lines

	Sequencing Output						Depth and Breadth of Coverage (Pf Genome)												Read Qualities		
Sample ID	Total bases (Gb)	Total reads	Mean read length (kb)	Median read length (kb)	Longest read length (kb)	Read length N50 (kb)	Mean depth of coverage	1x breadth of coverage	5x breadth of coverage	10x breadth of coverage	20x breadth of coverage	30x breadth of coverage	50x breadth of coverage	75x breadth of coverage	100x breadth of coverage	150x breadth of coverage	200x breadth of coverage	250x breadth of coverage	Proportion >Q5	Proportion >Q10	Proportion >Q15
XHA	3.42	388,561	8.80	4.58	261.32	14.63	120.30	99.38%	98.67%	98.21%	97.26%	96.51%	95.09%	92.88%	86.91%	5.38%	1.53%	0.95%	99%	76%	16%
BB12	2.07	298,286	6.95	3.44	157.32	15.26	76.06	99.55%	98.85%	98.08%	96.46%	94.63%	89.90%	54.85%	5.50%	1.54%	1.14%	0.70%	92%	69%	0%

Table S2B. Nanopore Sequencing Output for Mock Infections

	Treatment			Sequencing Output										Read Qualities			Read Alignment	
	Digestion prior to WGA	Purification after T7 endonuclease treatment	Clean-up after end prep	Total bases (Gb)	Total reads	Median read length (kb)	Mean read length (kb)	Longest read length (kb)	1x breadth of coverage	5x breadth of coverage	10x breadth of coverage	20x breadth of coverage	30x breadth of coverage	Proportion >Q5	Proportion >Q10	Proportion >Q15	Proportion mapping to Pf	Proportion mapping to Human
3D7 Mock Infection	McrBC digestion	2.5V ethanol precipitation	1V bead purification	0.06	25292	1.73	2.55	34.64	75.97%	12.17%	0.72%	0.00%	0.00%	100%	96%	19%	83%	17%
	McrBC digestion	2.5V ethanol precipitation	0.45V bead purification	0.26	81719	2.31	3.13	49.85	96.76%	76.09%	36.56%	4.02%	0.00%	100%	96%	19%	84%	16%
	McrBC digestion	1.8V bead purification	1V bead purification	0.12	48950	1.63	2.43	41.98	91.00%	37.69%	5.77%	0.11%	0.00%	100%	95%	21%	87%	13%
	McrBC digestion	1.8V bead purification	0.45V bead purification	0.25	80543	3.12	2.29	44.61	96.69%	76.54%	37.17%	3.84%	0.00%	100%	96%	21%	85%	15%
	Undigested	1.8V bead purification	1V bead purification	0.21	79412	2.68	1.74	62.94	27.88%	0.29%	0.00%	0.00%	0.00%	100%	92%	5%	5%	95%
	Undigested	1.8V bead purification	0.45V bead purification	0.28	87727	3.24	2.36	57.00	30.28%	0.34%	0.00%	0.00%	0.00%	100%	94%	5%	4%	96%

Table S2C. Nanopore Sequencing Output for Field Isolates

	Initial qPCR results	rWGA results				Sequencing Output						Depth and Breadth of Coverage (Pf Genome)						Read Qualities			Read Alignment				
Sample ID	Parasite density (copies/µL)	Av. parasite density (copies/µL)	St. dev. parasite density (copies/µL)	Av. fold enrichment	St. dev. fold enrichment	Total bases (Gb)	Total reads	Mean read length (kb)	Median read length (kb)	Longest read length (kb)	Read length N50 (kb)	Mean depth of coverage	1x breadth of coverage	5x breadth of coverage	10x breadth of coverage	20x breadth of coverage	30x breadth of coverage	Proportion >Q5	Proportion >Q10	Proportion >Q15	Proportion mapping to Pf	Proportion mapping to Pv	Proportion mapping to Human	Proportion unmapped	
PNG_16_1	27000	1.22E+10	6.79E+08	4.52E+05	2.51E+04	0.26	90,984	2.89	2.03	39.37	4.20	9.61	96.00%	75.80%	43.25%	8.52%	1.38%	100%	94%	20%	85%	0%	3%	12%	SELECT FOR SEQUENCING
PNG_12_1a	19500	1.03E+09	2.94E+08	5.29E+04	1.51E+04	0.10	40,235	2.60	1.83	54.75	3.72	2.32	76.10%	15.28%	1.29%	0.02%	0.00%	98%	17%	0%	52%	0%	35%	13%	
PNG_12_1	20864	7.62E+08	2.38E+07	3.65E+04	1.14E+03	0.87	447,858	1.95	1.44	48.52	2.75	16.43	94.27%	80.33%	64.52%	32.67%	15.36%	100%	39%	0%	44%	0%	40%	16%	
PNG_05_5	21000	4.98E+08	1.99E+08	2.37E+04	9.46E+03	2.23	1,176,980	1.89	1.41	155.47	2.54	30.01	97.67%	91.37%	81.44%	58.64%	38.76%	97%	10%	0%	31%	0%	52%	17%	
PNG_16_4	22800	4.84E+08	2.52E+08	2.12E+04	1.11E+04	0.05	13,435	4.00	2.72	83.04	5.46	1.11	52.22%	3.83%	0.11%	0.00%	0.00%	100%	54%	0%	48%	0%	40%	12%	
PNG_16_6	587	4.76E+08	4.24E+08	8.10E+05	7.22E+05	0.34	132,178	2.58	1.65	97.67	17.94	6.14	92.03%	55.81%	20.95%	1.84%	0.17%	99%	16%	0%	42%	0%	42%	16%	
PNG_05_4	16840	4.27E+08	3.74E+07	2.54E+04	5.58E+03	0.40	211,320	1.89	1.35	37.15	2.68	3.87	86.03%	33.14%	7.18%	0.40%	0.06%	98%	14%	0%	23%	0%	65%	13%	
PNG_16_7	1860	4.01E+08	4.03E+08	2.15E+05	2.17E+05	0.51	165,355	3.09	2.04	69.42	13.95	9.19	95.53%	73.76%	41.19%	7.39%	1.01%	98%	16%	0%	42%	0%	44%	14%	
PNG_16_3	13900	3.41E+08	2.96E+08	2.45E+04	2.13E+04	0.30	73,487	4.12	2.79	93.60	6.42	2.28	76.72%	14.72%	1.05%	0.00%	0.00%	100%	50%	0%	18%	0%	71%	11%	
PNG_16_5	1920	3.20E+08	3.60E+08	1.66E+05	1.88E+05	0.18	54,699	3.33	2.17	62.73	9.62	2.81	80.32%	21.41%	2.46%	0.04%	0.00%	100%	49%	0%	36%	0%	51%	13%	
PNG_05_3	20920	3.19E+08	2.96E+08	1.53E+04	1.42E+04	0.97	464,915	2.08	1.46	44.49	2.95	7.56	95.44%	65.39%	28.92%	4.17%	0.65%	100%	62%	0%	14%	0%	52%	34%	
PNG_16_2	32200	1.59E+08	1.53E+08	4.94E+03	4.77E+03	0.69	238,264	2.90	1.99	52.46	4.24	21.54	97.76%	91.24%	79.40%	47.46%	23.67%	100%	93%	17%	73%	0%	15%	13%	
PNG_05_7	12080	1.52E+08	7.48E+07	1.26E+04	6.19E+03	1.15	616,664	1.86	1.32	42.86	2.70	6.48	81.70%	48.82%	26.05%	4.67%	0.83%	100%	36%	0%	13%	0%	71%	16%	
PNG_05_6	21120	1.39E+08	4.13E+07	6.57E+03	1.96E+03	2.64	1,487,301	1.77	1.30	87.14	2.38	3.92	85.34%	33.28%	8.08%	0.58%	0.09%	98%	7%	0%	3%	0%	79%	18%	
PNG_05_2	12700	1.35E+08	1.21E+08	1.06E+04	9.49E+03	1.03	533,841	1.93	1.35	38.88	2.70	4.01	82.21%	32.53%	9.64%	1.20%	0.25%	100%	59%	0%	9%	0%	75%	16%	
PNG_16_9	15300	1.34E+08	1.95E+08	8.73E+03	1.27E+04	0.07	26,400	2.78	1.79	62.06	1.75	0.76	43.06%	1.60%	0.03%	0.00%	0.00%	99%	19%	0%	24%	0%	62%	14%	
PNG_05_8	1556	1.32E+08	1.50E+08	8.49E+04	9.66E+04	1.26	670,554	1.88	1.34	36.14	2.72	5.48	74.65%	39.97%	20.37%	4.10%	0.98%	100%	36%	0%	10%	0%	74%	16%	
PNG_12_2	7050	5.60E+07	3.26E+07	7.94E+03	4.63E+03	0.08	27,501	2.75	1.87	60.39	4.12	0.19	14.05%	0.09%	0.00%	0.00%	0.00%	98%	9%	0%	6%	0%	82%	12%	DO NOT SEQUENCE
PNG_05_9	2084	4.98E+07	1.50E+07	2.39E+04	7.18E+03	1.89	971,238	1.95	1.44	70.79	2.62	1.45	40.61%	7.77%	1.55%	0.05%	0.00%	100%	33%	0%	2%	0%	81%	17%	
PNG_05_10	8160	4.83E+07	4.52E+07	5.92E+03	5.54E+03	0.19	67,094	2.70	1.64	57.29	4.50	1.40	57.59%	6.90%	0.47%	0.00%	0.00%	100%	45%	2%	18%	0%	63%	20%	
Cam_01	960	2.09E+07	2.22E+07	2.18E+04	2.32E+04	0.25	125,243	2.03	1.40	37.54	1.11	3.23	68.47%	22.70%	8.00%	1.79%	0.58%	100%	44%	0%	30%	0%	35%	36%	
PNG_05_1	12560	1.94E+07	5.98E+06	1.55E+03	4.76E+02	0.57	339,590	1.68	1.21	34.30	2.42	4.27	72.59%	28.27%	10.90%	2.96%	1.25%	100%	31%	0%	17%	0%	59%	23%	
Cam_02	892	1.64E+07	9.99E+06	1.45E+04	1.68E+04	0.14	69,522	2.07	1.45	45.49	1.86	1.36	51.30%	7.36%	1.19%	0.10%	0.02%	100%	44%	0%	22%	0%	36%	43%	
PNG_16_8	2520	1.41E+07	2.12E+07	5.61E+03	8.40E+03	0.21	72,300	2.86	1.83	59.95	4.58	4.93	90.28%	46.51%	12.25%	0.58%	0.02%	100%	53%	0%	56%	0%	31%	13%	
PNG_05_11	916	8.91E+06	5.92E+06	9.73E+03	6.46E+03	0.51	294,020	1.75	1.23	49.38	2.57	0.32	21.54%	0.39%	0.02%	0.00%	0.00%	100%	34%	0%	1%	0%	81%	18%	
PNG_05_12	3000	1.05E+06	5.10E+05	3.49E+02	1.70E+02	0.69	384,101	1.78	1.78	55.98	2.62	0.19	13.70%	0.20%	0.01%	0.00%	0.00%	100%	66%	0%	1%	0%	86%	13%	
PNG_05_13	2340	2.66E+05	2.84E+05	1.14E+02	1.22E+02	0.27	96,700	2.78	1.78	62.29	4.44	0.73	39.19%	1.83%	0.08%	0.00%	0.00%	100%	48%	0%	6%	0%	79%	15%	
PNG_16_10	334	5.22E+06	3.45E+06	1.56E+04	1.03E+04	0.09	37,941	2.26	1.55	36.73	3.29	0.09	6.64%	0.05%	0.00%	0.00%	0.00%	98%	8%	0%	2%	0%	77%	20%	DO NOT ENRICH
PNG_05_14	276	1.75E+07	1.11E+07	6.35E+04	4.02E+04	0.46	328,600	1.41	0.99	29.91	2.09	2.17	63.13%	14.95%	3.34%	0.34%	0.07%	100%	35%	0%	11%	0%	68%	21%	
Cam_03	68	4.56E+06	5.21E+06	6.70E+04	7.66E+04	0.19	78,529	2.38	1.64	64.95	9.47	1.50	29.24%	8.29%	3.90%	1.38%	0.62%	100%	49%	0%	19%	0%	51%	31%	
PNG_12_3	23	1.56E+07	1.07E+07	6.77E+05	4.66E+05	1.34	671,992	1.99	1.38	64.53	3.01	0.70	41.67%	1.07%	0.02%	0.00%	0.00%	100%	33%	0%	1%	0%	86%	13%	
PNG_12_4	2.59	1.93E+04	3.12E+04	7.46E+03	1.20E+04	0.82	402,397	2.03	1.36	59.47	3.20	0.01	0.78%	0.02%	0.01%	0.00%	0.00%	100%	32%	0%	0%	0%	87%	13%	
PNG_12_5	0.674	1.67E+04	2.89E+04	2.48E+04	4.28E+04	0.99	523,996	1.90	1.32	65.81	2.81	0.12	8.01%	0.13%	0.01%	0.00%	0.00%	100%	32%	0%	0%	0%	85%	15%	

Table S3. SNV Genotyping Concordance Rates in Homopolymeric Regions

Sample: XHA (106X coverage)

	Genotype	*Unfiltered*	*Filtered*
Homopolymer <=6bp	Concordant	683731	676846
	Discordant	1649	507
Homopolymer >6bp	Concordant	7823	7641
	Discordant	106	19
Discordance in homopolymers <=6bp:		0.24%	0.08%
Discordance in homopolymers >6bp		1.34%	0.25%

Sample: BB12 (75X coverage)

	Genotype	*Unfiltered*	*Filtered*
Homopolymer <=6bp	Concordant	656892	649945
	Discordant	1179	399
Homopolymer >6bp	Concordant	7560	7475
	Discordant	99	54
Discordance in homopolymers <=6bp:		0.18%	0.06%
Discordance in homopolymers >6bp		1.29%	0.72%

Sample: Cam_01 (3.2X coverage)

	Genotype	*Unfiltered*	*Filtered*
Homopolymer <=6bp	Concordant	451550	303559
	Discordant	3528	529
Homopolymer >6bp	Concordant	5076	3170
	Discordant	161	27
Discordance in homopolymers <=6bp:		0.78%	0.17%
Discordance in homopolymers >6bp		3.07%	0.84%

Sample: Cam_02 (1.4X coverage)

	Genotype	*Unfiltered*	*Filtered*
Homopolymer <=6bp	Concordant	329574	169799
	Discordant	3083	282
Homopolymer >6bp	Concordant	3628	1753
	Discordant	146	23
Discordance in homopolymers <=6bp:		0.93%	0.17%
Discordance in homopolymers >6bp		3.87%	1.30%