An In-Solution Hybridisation Method for the Isolation of Pathogen DNA from Human DNA-rich Clinical Samples for Analysis by NGS

Miriam Smith[1], Susana Campino[1], Yong Gu[1], Taane G. Clark[2], Thomas D. Otto[1], Gareth Maslen[1], Magnus Manske[1], Mallika Imwong[3], Arjen M. Dondorp[4,5], Dominic P. Kwiatkowski[1,6], Michael A. Quail[1,*] and Harold Swerdlow[1]

[1]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK*

[2]*London School of Hygiene and Tropical Medicine, Keppel Street, London, UK*

[3]*Department of Molecular Tropical Medicine and Genetics, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand*

[4]*Mahidol Oxford Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand*

[5]*Centre for Tropical Medicine, Churchill Hospital, Oxford, UK*

[6]*Wellcome Trust Centre of Human Genetics, University of Oxford, Oxford, UK*

Abstract: Studies on DNA from pathogenic organisms, within clinical samples, are often complicated by the presence of large amounts of host, e.g., human DNA. Isolation of pathogen DNA from these samples would improve the efficiency of next-generation sequencing (NGS) and pathogen identification. Here we describe a solution-based hybridisation method for isolation of pathogen DNA from a mixed population. This straightforward and inexpensive technique uses probes made from whole-genome DNA and off-the-shelf reagents.

In this study, *Escherichia coli* DNA was successfully enriched from a mixture of *E.coli* and human DNA. After enrichment, genome coverage following NGS was significantly higher and the evenness of coverage and GC content were unaffected. This technique was also applied to samples containing a mixture of human and *Plasmodium falciparum* DNA. The *P.falciparum* genome is particularly difficult to sequence due to its high AT content (80.6%) and repetitive nature. Post enrichment, a bias in the recovered DNA was observed, with a poorer representation of the AT-rich non-coding regions. This uneven coverage was also observed in pre-enrichment samples, but to a lesser degree. Despite the coverage bias in enriched samples, SNP (single-nucleotide polymorphism) calling in coding regions was unaffected and the majority of samples had over 90% of their coding region covered at 5x depth.

This technique shows significant promise as an effective method to enrich pathogen DNA from samples with heavy human contamination, particularly when applied to GC-neutral genomes.

Keywords: AT-rich DNA, clinical samples, *E.coli*, enrichment, host DNA contamination, in-solution hybridisation, next generation sequencing, *P.falciparum*.

INTRODUCTION

According to the World Health Organisation, infectious diseases are the world's biggest killer of children and young people, causing 15 million deaths per year. If the causal pathogen can be identified, many of these diseases are treatable via vaccination and/or drug-treatment regimens. Point-of-care microbial or pathoclinical DNA sample collection for the identification of pathogens can be difficult, particularly in the developing world and during pathogenic outbreaks. This means that clinical samples are often composed of a mixture of pathogen and host DNA. For example, when blood samples are taken for the detection of the malaria parasite *Plasmodium*, there is often significant contamination with human DNA. Although next-generation sequencing has the capacity to sequence all of the DNA contained within a clinical sample, this is very inefficient and increases the cost, computational burden and duration of the analyses. In addition, the human component is often preferentially sequenced because the methodologies are optimised for human DNA and this genome has a relatively neutral GC content. This neutrality makes the amplification and sequencing easier than for many pathogens that have extreme GC contents, for example the *Plasmodium* genome contains less than 20% GC bases [1].

*Address correspondence to this author at the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK, CB10 1SA; Tel: +44 (0)1223 494819; Fax: +44 (0)1223 494919;
E-mail: mq1@sanger.ac.uk

1875-693X/12

It would be beneficial to be able to isolate the pathogen DNA from the unwanted human contamination. This would ensure that only the DNA of interest is being sequenced, reducing the amount of sequencing required to obtain adequate coverage and depth and enable meaningful analysis of the data. The majority of studies rely upon growing pathogens in the laboratory, which is a time consuming approach, is not always successful and is subject to potential selective pressures. Commercial kits for the isolation of bacterial DNA from whole blood are available. The MolYsis kit from Molzym (www.molzym.com) works by lysing human cells under chaotropic conditions which do not affect bacterial cells due to their more rigid cell walls [2, 3]. The LOOXSTER™ method commercialised by SIRS-Lab (www.sirs-lab.de) isolates bacterial DNA based upon the presence of non-methylated CpG dinucleotides in the bacterial but not in the human DNA [2, 3]. Neither of these techniques will work for all pathogen types, because not all types of microbes have rigid cell walls or contain unmethylated CpG dinucleotides. Another method for separation of pathogen and host DNA utilises Hoechst dye and ultracentrifugation in a CsCl gradient; however, this will only isolate DNA with a significantly different GC content to the host [4]. The Hoechst dye method can be useful for the isolation of *Plasmodium* DNA, but has limited utility for other pathogens [4]. In addition, it requires large amounts of starting material and is both inefficient and low-throughput. Isolation of microbial cells from blood samples has also been carried out using various density gradients and magnetic isolation techniques [5-7].

For sequencing-based studies, another approach to selectively isolate DNA would use oligonucleotide probes designed specifically against the pathogen under investigation. Many target-enrichment methods for next-generation sequencing applications have been developed, including those commercialised by Agilent Technologies (www.genomics.agilent.com), Roche Nimblegen (www.nimblegen.com), RainDance Technologies (www.raindance technologies.com), Halo Genomics (www.halogenomics.com) and Illumina (www.illumina.com) [8-11]. The most widely used technique is Agilent's SureSelect Target Enrichment System, which is a hybridisation-based sequence capture using a set of oligonucleotide capture probes, either in-solution or attached to a solid support [9]. Currently, the main use of these target-enrichment technologies is for human exome sequencing and targeted selection of regions of the human genome. However, it would also be possible to purchase a library of probes designed against any small genome of interest which could then be used to enrich clinical samples for that organism and help to eliminate human contamination. However, even for the isolation of small genomes this would be very costly.

In order to reduce the cost of target-enrichment techniques for the isolation of microbial DNA from clinical samples, we have developed a whole-genome pathogen capture technique. This technique is a solution-based target enrichment approach, similar to that used by Agilent, which utilises off-the-shelf reagents and home-made probes which can be made using whole-genome DNA from any organism of interest, exemplified in this study by *E.coli* or *P.falciparum* [9]. In this study, probes were made by fragmentation and biotinylation of whole-genome DNA from the organism of interest. A similar method was recently published by Melnikov *et al.,* in Genome Biology [12] where home-made RNA baits were used. The DNA samples under investigation in our study were subjected to Illumina library preparation before denaturation at high temperature of both the library and probes and subsequent hybridisation together for 24 hours before isolation of the captured material using streptavidin-coated beads. The isolated DNA was then quantified by qPCR (quantitative polymerase chain reaction) and sequenced using next-generation sequencing technology (Illumina). As an alternative strategy, whole-genome human probes were produced utilising the same protocol and these were used to remove the human contamination from a mixed human/pathogen sample. A similar hybridisation technique to that described above was employed, collecting and sequencing the hybridisation eluate rather than the probe-bound DNA.

The solution capture method developed in this study has applications for the identification of pathogens in clinical samples. Index tagging (multiplexing) of DNA samples would enable several patient samples to be hybridised and sequenced simultaneously. Alternatively, it should be possible to selectively target multiple pathogens in the same capture experiment, for example, for detection from metagenomic samples. Whole-genome solution capture has many potential clinical applications, as it offers a cost-effective method to isolate pathogens from clinical samples. This allows the whole-genome sequencing of the isolated pathogens, opening up new opportunities in infectious-disease control and prevention.

In this study, a whole-genome, solution-based hybridisation technique has been developed that can be used to isolate pathogen DNA from samples with high levels of human contamination. This technique was tested with *E.coli* and *P.falciparum* and was shown to provide enrichment of pathogen DNA in each case. This has the potential for clinical utility, especially for GC-neutral genomes.

RESULTS AND DISCUSSION

E.coli

Preliminary experiments were carried out to isolate *E.coli* DNA from mixed samples of *E.coli* and human DNA. *E.coli* was chosen because of the small size (~4.6Mb) and neutral GC content (50.8% GC) of its genome [13]. A range of mixtures were attempted including 10:90, 25:75 and 50:50 *E.coli*:human DNA. In all cases, the proportion of *E.coli* DNA was significantly enriched after the pulldown procedure, with over 85% of the sequencing reads aligning to the *E.coli* genome after pulldown enrichment (Table **1**, **1A**). The depth of coverage throughout the genome was significantly higher after enrichment, further demonstrating the utility of this technique (Table **2**, **1A**). The GC content and the evenness of coverage were comparable between the pre- and post-enrichment samples, showing that the enrichment procedure is not biased for this genome (Figs. **1A** and **1C**).

Further experiments were carried out to determine the optimum DNA requirements for probe synthesis and input

Table 1. Enrichment Statistics for the *E.coli* Genome Pre- and Post- solution-based Pulldown

Experiment	Sample ID	Sample Name	Probe	Read Count	% *E.coli*	% Human	% Other
1A	1A i	*E.coli*:human 50:50 prepulldown	~	10311001	36.04	26.66	37.30
	1A ii	*E.coli*:human 25:75 prepulldown	~	13668667	28.25	59.71	12.05
	1A iii	*E.coli*:human 10:90 prepulldown	~	17347896	10.53	69.99	19.48
	1A iv	*E.coli*:human 50:50 pulldown	*E.coli* 1.5µg	15067104	87.49	6.86	5.65
	1A v	*E.coli*:human 25:75 pulldown	*E.coli* 1.5µg	11206843	88.33	5.04	6.63
	1A vi	*E.coli*:human 10:90 pulldown	*E.coli* 1.5µg	11884992	85.03	9.08	5.89
1B	1B i	*E.coli*:human 20:80 prepulldown	~	12757413	9.30	38.32	52.37
	1B ii	*E.coli*:human 20:80 800ng	*E.coli* 3µg	20079543	79.85	5.31	14.84
	1B iii	*E.coli*:human 20:80 1600ng	*E.coli* 3µg	25326464	76.41	7.71	15.88
	1B iv	*E.coli*:human 20:80 prepulldown	~	38139698	9.47	55.42	35.11
	1B v	*E.coli*:human 20:80 800ng	*E.coli* 1.5µg	45334592	80.41	7.98	11.61
3A	3A i	*E.coli:*human 20:80 prepulldown	~	38139698	9.47	55.42	35.11
	3A ii	*E.coli:*human 20:80	*E.coli* 1.5µg	45334592	80.41	7.98	11.61
	3A iii	*E.coli:*human 20:80	human 10µg	38158190	4.47	72.26	23.27
	3A iv	*E.coli:*human 20:80 eluate	human 10µg	34334647	11.06	45.87	43.08
	3A v	*E.coli:*human 20:80 eluate	human 3µg	33342685	10.51	49.07	40.42
	3A vi	*E.coli:*human 20:80 *E.coli* probe eluate	*E.coli* 1.5µg	42189483	6.24	61.93	31.83

Probe amount refers to the amount of DNA used at the start of probe synthesis. Where an amount of sample is listed this is the total amount of DNA added to the hybridisation reaction. Read count is the number of reads obtained per single lane of paired-end Illumina GA sequencing. The proportion of the reads mapping to *E.coli* and human genomes are listed. The 'Other' category includes all reads that do not map to either the *E.coli* or human genomes and will include adapter dimers.

Table 2. Genome Coverage Statistics for the *E.coli* Genome Pre- and Post- solution-based Pulldown

Experiment	Sample ID	Sample Name	Probe	Average Depth	Maximum Depth
1A	1A i	*E.coli*:human 50:50 prepulldown	~	63	136
	1A ii	*E.coli*:human 25:75 prepulldown	~	65	140
	1A iii	*E.coli*:human 10:90 prepulldown	~	31	78
	1A iv	*E.coli*:human 50:50	*E.coli* 1.5µg	221	656
	1A v	*E.coli*:human 25:75	*E.coli* 1.5µg	166	512
	1A vi	*E.coli*:human 10:90	*E.coli* 1.5µg	170	501
1B	1B i	*E.coli*:human 20:80 prepulldown	~	41	94
	1B ii	*E.coli*:human 20:80 800ng	*E.coli* 3µg	555	1565
	1B iii	*E.coli*:human 20:80 1600ng	*E.coli* 3µg	669	1912
			~		
	1B iv	*E.coli*:human 20:80 prepulldown	*E.coli* 1.5µg	126	288
	1B v	*E.coli*:human 20:80 800ng	*E.coli* 1.5µg	1267	4398

Average depth refers to the mean number of Illumina GA sequencing reads that can be mapped to each position across the genome and the maximum depth refers to the greatest number of reads mapped to a single location in the genome. Probe amount refers to the amount of DNA used at the start of probe synthesis. Where an amount of sample is listed this is the total amount of DNA added to the hybridisation reaction.

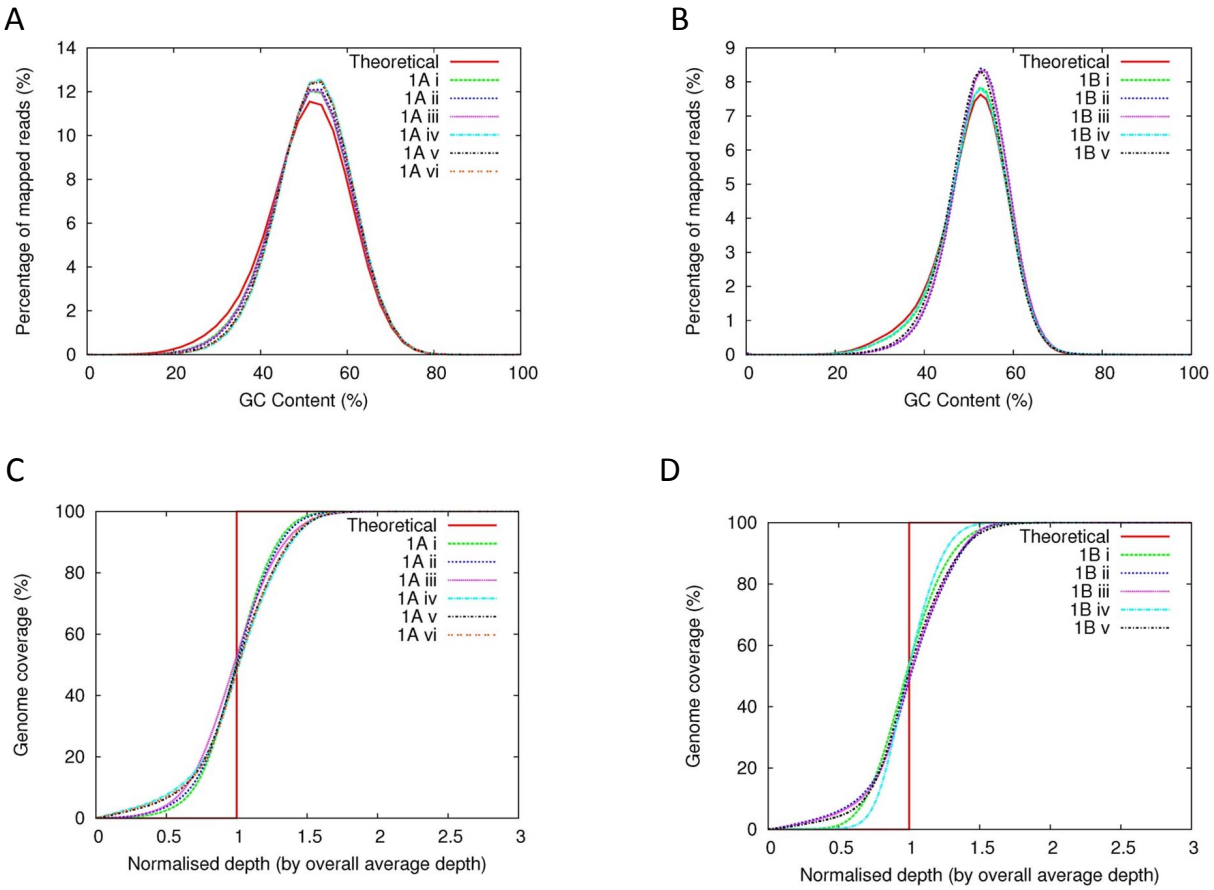

Fig. (1). GC content and genome coverage distribution of *E.coli* samples pre- and post- solution-based pulldown. **A**) The percentage of mapped sequencing reads distributed by their GC content for; 1A i - *E.coli*:human 50:50 prepulldown, 1A ii - *E.coli*:human 25:75 prepulldown, 1A iii - *E.coli*:human 10:90 prepulldown, 1A iv - *E.coli*:human 50:50, 1A v - *E.coli*:human 25:75, 1A vi - *E.coli*:human 10:90. The red line shows the theoretical value for the reference genome. **B**) The percentage of mapped sequencing reads distributed by their GC content for; 1B i - *E.coli*:human 20:80 prepulldown, 1B ii - *E.coli*:human 20:80 800ng 3µg probe, 1B iii - *E.coli*:human 20:80 1600ng 3µg probe, 1B iv - *E.coli*:human 20:80 prepulldown, 1B v - *E.coli*:human 80:20 800ng 1.5µg probe. The red line shows the theoretical normalised read depth across the *E.coli* genome. The sample curves show how much coverage is above and below the average coverage level. The further the sample curve is away from the theoretical line, the more biased the coverage. **C**) Genome coverage distribution plots for; 1A i - *E.coli*:human 50:50 prepulldown, 1A ii - *E.coli*:human 25:75 prepulldown, 1A iii - *E.coli*:human 10:90 prepulldown, 1A iv - *E.coli*:human 50:50, 1A v - *E.coli*:human 25:75, 1A vi - *E.coli*:human 10:90. The red line shows the theoretical normalised read depth across the *E.coli* genome. The sample curves show how much coverage is above and below the average coverage level. The further the sample curve is away from the theoretical line, the more biased the coverage. **D**) Genome coverage distribution plots for; 1B i - *E.coli*:human 20:80 prepulldown, 1B ii - *E.coli*:human 20:80 800ng 3µg probe, 1B iii - *E.coli*:human 20:80 1600ng 3µg probe, 1B iv - *E.coli*:human 20:80 prepulldown, 1B v - *E.coli*:human 80:20 800ng 1.5µg probe. Commentary as in part C.

DNA. The conditions evaluated were 1600ng or 800ng of library for enrichment and 1.5µg or 3µg starting DNA for probe synthesis. There was little difference between the conditions under investigation in terms of enrichment or GC content (Fig. **1B** and Table **1**, **1B**). The increase in coverage depth after enrichment was similar for the different conditions and the evenness of coverage was also similar (Fig. **1D** and Table **2**, **1B**). 800ng of input DNA and 1.5µg of probe (initial amount for probe synthesis) were chosen for further experiments. During the development of this technique, three different biotinylation techniques were trialled for the synthesis of the probe molecules and it was found that they all provided similar results in terms of enrichment and sequencing statistics (**see Supplementary Data Files 1 - 5**).

P.falciparum

The *Plasmodium* genome has been a strong research focus of the Wellcome Trust Sanger Institute through its involvement in the sequencing of the reference strain 3D7, and subsequent genomic diversity studies in laboratory and field isolates (Sanger Institute Malarial Programme www.sanger.ac.uk/malaria) because of the high disease burden inflicted by malaria. *Plasmodium* was predicted to be a more challenging organism for solution-based capture due to its larger genome (22.8Mb) and high content of AT bases

(80.6%) [1]. In this study, initial experiments were carried out using probes made from the *P.falciparum* laboratory strain 3D7 and these were used to enrich for the same strain within test mixtures of *Plasmodium* and human DNA (25:75 *Plasmodium*:human). Different amounts of probe (1.5µg and 3µg of starting DNA for probe synthesis) and library (500ng, 800ng and 1600ng) were assessed to see which provided the best combination for enrichment (Table **3**, **2A** and **2B**). As previously seen with the experiments using the *E.coli* probe and DNA, a high level of enrichment was seen with all samples (on average 71.3% of sequencing reads aligning to the *Plasmodium* genome after enrichment) and the best results were obtained with 800ng of library and 1.5µg of starting material used for probe synthesis (Table **3**, **2A** and **2B**). Despite the high levels of enrichment, a comparison of the GC content of the pre- and post-capture samples showed that the enrichment did not work well for AT-rich regions and the GC content of the post-enrichment samples was higher than the average for both the theoretical reference genome (19.4%) and the actual sequencing of the pre-pulldown samples (Fig. **2A** and **2B**). This is likely to be caused by differences in the hybridisation temperature and efficiency of AT- and GC-rich DNA sequences [14, 15].

As an initial measure to counteract this enrichment bias, the hybridisation temperature was lowered from 65°C to 55°C. A lower hybridisation temperature was hypothesised to encourage better annealing of AT-rich sequences [14, 15]. However, in these experiments, the hybridisation temperature appeared to have little effect on the GC content of the enrichment and the lower hybridisation temperature actually reduced the overall enrichment of *Plasmodium* DNA

Table 3. Enrichment Statistics for the *P.falciparum* Genome Pre- and Post- solution-based Pulldown

Experiment	Sample ID	Sample Name	Probe Strain	Read Count	% *P.falciparum*	% Human	% Other
2A	2A i	3D7:human 25:75 prepulldown	~	37261895	14.97	52.88	32.15
	2A ii	3D7:human 25:75 800ng	3D7 1.5µg	36139494	72.44	8.22	19.34
	2A iii	3D7:human 25:75 500ng	3D7 1.5µg	32434631	72.75	7.84	19.41
2B	2B i	3D7:human 25:75 prepulldown	~	40553893	13.42	46.92	39.67
	2B ii	3D7:human 25:75 1600ng	3D7 1.5µg	21241885	66.41	15.90	17.68
	2B iii	3D7:human 25:75 1600ng	3D7 3µg	28703869	68.34	15.34	16.32
	2B iv	3D7:human 25:75 800ng	3D7 1.5µg	41465650	74.77	10.55	14.68
	2B vi	3D7:human 25:75 800ng	3D7 3µg	29455966	73.19	12.19	14.61
2C	2C i	3D7:human 30:70 prepulldown	~	26938319	13.27	46.85	39.88
	2C ii	3D7:human 30:70 55°C hyb	3D7 1.5µg	7909731	67.95	10.29	21.75
	2C iii	3D7:human 30:70 65°C hyb	3D7 1.5µg	8068594	75.64	5.90	18.46
2D	2D i	3D7:human 20:80 prepulldown	~	35442819	11.85	43.56	44.59
	2D ii	3D7:human 20:80	3D7 1.5µg	44006702	74.34	10.21	15.45
	2D iii	3D7:human 20:80 1M TMAC	3D7 1.5µg	38447345	76.82	9.04	14.14
	2D iv	3D7:human 20:80 0.5M TMAC	3D7 1.5µg	44090067	75.20	10.34	14.46
	2D vi	3D7:human 20:80 0.1M TMAC	3D7 1.5µg	44302101	77.08	8.69	14.22
2E	2E i	Dd2:human 30:70 prepulldown	~	41578455	27.06	28.27	44.67
	2E ii	Dd2:human 30:70	3D7 1.5µg	33817623	67.49	4.15	28.36
2F	2F i	3D7:human 20:80 prepulldown	~	25542843	9.01	41.08	49.91
	2F ii	3D7:human 20:80	3D7, Dd2, IT and HB3	34060230	75.21	7.09	17.69
2G	2G i	Isolate 1:human 20:80 prepulldown	~	45566177	8.88	56.15	34.98
	2G ii	Isolate 2:human 20:80 prepulldown	~	37065625	6.41	51.93	41.66
	2G iii	Isolate 1:human 20:80 mix probe	3D7, Dd2, IT and HB3	36955463	67.19	12.44	20.36
	2G iv	Isolate 2:human 20:80 mix probe	3D7, Dd2, IT and HB3	18331991	65.47	14.89	19.64

Enrichment determined by comparison of Illumina sequencing reads to *P.falciparum* 3D7 and human reference genomes. Probe amount refers to the amount of DNA used at the start of probe synthesis. Where an amount of sample is listed this is the total amount of DNA added to the hybridisation reaction. Read count is the number of reads obtained per single lane of Illumina GA sequencing. The proportion of the reads mapping to *P.falciparum* and human genomes are listed. The 'Other' category includes all reads that do not map to either the *P.falciparum* or human genomes and will include primer dimers.

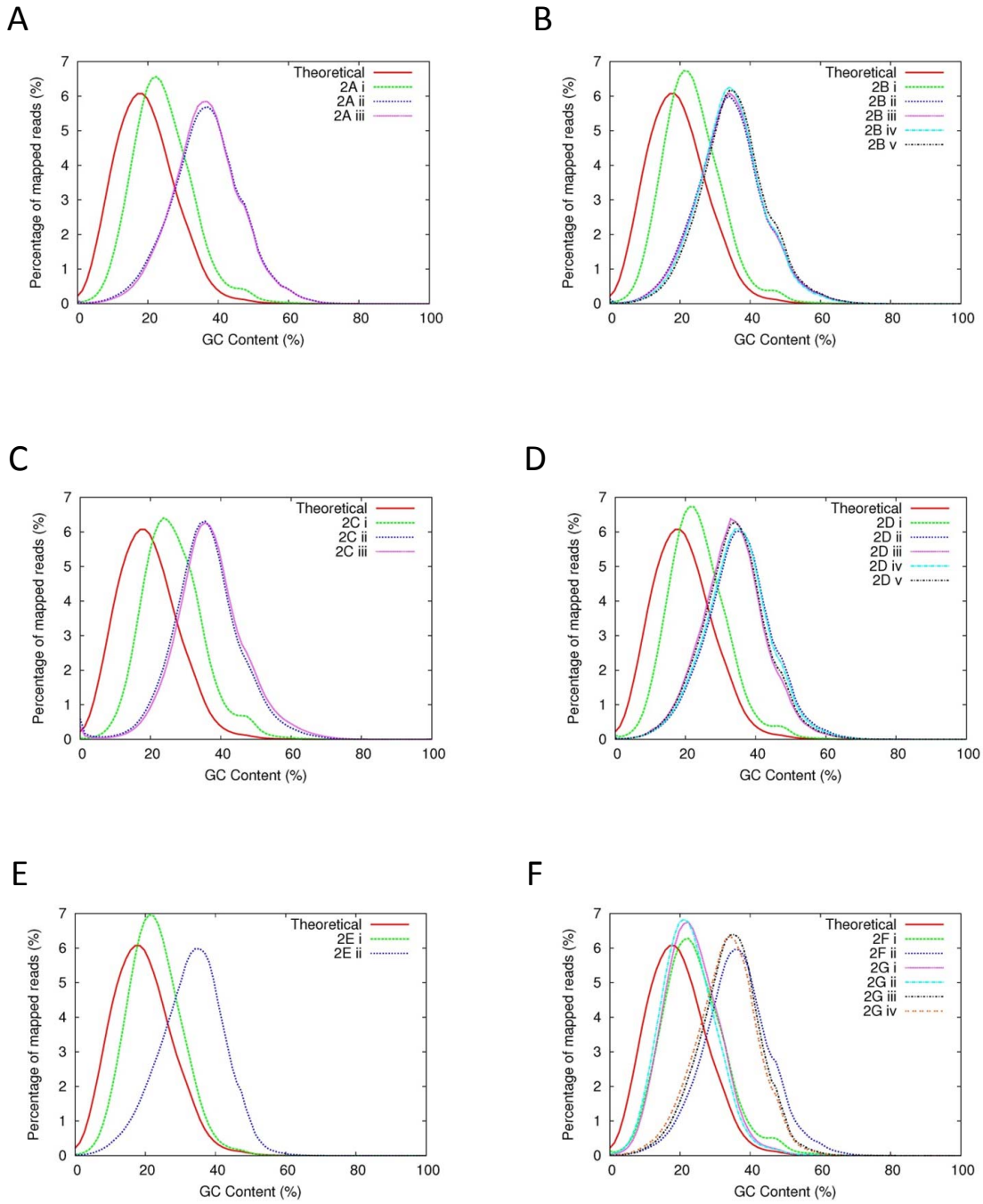

Fig. (2). GC content of *P.falciparum* samples pre- and post- solution-based pulldown. The percentage of mapped sequencing reads distributed by their GC content. The red line shows the theoretical value for the 3D7 reference genome. **A**) 2A i - 3D7:human 25:75 prepulldown, 2A ii - 3D7:human 25:75 800ng, 1.5µg probe, 2A iii - 3D7:human 25:75 500ng, 1.5µg probe. **B**) 2B i - 3D7:human 25:75 prepulldown, 2B ii - 3D7:human 25:75 1600ng, 1.5µg probe, 2B iii - 3D7:human 25:75 1600ng, 3µg probe, 2B iv - 3D7:human 25:75 800ng, 1.5µg probe, 2B v - 3D7:human 25:75 800ng, 3µg probe. **C**) 2C i - 3D7:human 30:70 prepulldown, 2C ii - 3D7:human 30:70 55^{o}C hyb, 1.5µg probe, 2C iii - 3D7:human 30:70 65^{o}C hyb, 1.5µg probe. **D**) 2D i - 3D7:human 20:80 prepulldown, 2D ii - 3D7:human 20:80, 1.5µg probe, 2D iii - 3D7:human 20:80 1M TMAC, 1.5µg probe, 2D iv - 3D7:human 20:80 0.5M TMAC, 1.5µg probe, 2D v - 3D7:human 20:80 0.1M TMAC, 1.5µg probe. **E**) 2E i - Dd2:human 30:70 prepulldown, 2E ii - Dd2:human 30:70, 1.5µg probe. **F**) 2F i - 3D7:human 20:80 prepulldown, 2F ii - 3D7:human 20:80 mixed probe, 2G i - Isolate 1:human 20:80 prepulldown, 2G ii - Isolate 2:human 20:80 prepulldown, 2G iii - Isolate 1:human 20:80 mixed probe, 2G iv - Isolate 2:human 20:80 mixed probe.

from the human mix by a small amount, possibly due to increased non-specific binding of the human DNA (Fig. **2C** and Table **3**, **2C**).

As an alternative method to counteract the high GC bias of the *Plasmodium* genome observed during sequencing after solution-based hybridisation, TMAC (tetramethyl

ammonium chloride, Sigma-Aldrich) was added to the hybridisation buffer. TMAC is an additive which stabilises AT base pairs, causing AT-interactions to behave similarly to GC-interactions and therefore acting to reduce bias effects across the genome [16, 17]. It has previously been shown that in 3M TMAC an AT base pair is as thermally stable as a GC base pair [17]. In these experiments, between 0.1M and 1M TMAC was added to the hybridisation buffer to increase the hybridisation of highly AT-rich regions of the *Plasmodium* genome. This additive had little effect on the amount of enrichment seen after pulldown and did not appear to significantly improve the hybridisation of the AT-rich regions (Fig. **2D** and Table **3**, **2D**). Possibly, insufficient TMAC was added to have a significant effect, however further TMAC could not be added due to the composition of the hybridisation buffer being such that it was impossible to add further TMAC to the buffer.

The extremely AT-rich regions of the *Plasmodium* genome are in non-coding areas of the genome and therefore the GC content of the exonic region (~23.7% GC) is not as extreme as seen for the non-coding regions (~13.5%) [1]. The exonic region is also the most interesting part of the genome as it contains all of the protein-coding information. A further analysis, concentrating on the enrichment of the coding regions after pulldown, demonstrated that the average and maximum coverage of the coding region were greatly increased after the enrichment procedure and up to 98.2% of the coding region was covered at 5x depth (Fig. **3**, Table **4**, **2A-2D**). The maximum depth of coverage was significantly higher after enrichment due to a lack of uniformity in coverage after pulldown has occurred (Fig. **3**, Table **4**, **2A-2D**). This is likely to be caused by the differences in hybridisation of GC- and AT-rich regions. Hybridisation temperature and TMAC addition did not significantly affect the amount or the eveness of coverage of the coding region recovered after hybridisation (Table **4**, **2C** and **2D**). Despite this coverage nonuniformity it was still possible to identify single nucleotide polymorphisms (SNPs) at 10 x coverage at a similar rate pre- and post-enrichment (Table **4**, **2A-2D**). This observation would suggest that the uneven coverage does not affect SNP calling and that there is sufficient coverage of the exome after enrichment for useful analytical measurements to be made.

In order to further assess the utility of this technique, it was important to test other strains of *P.falciparum*. In a preliminary experiment, probes were made from *P.falciparum* laboratory strain 3D7 and these probes were used to selectively hybridise DNA from *P.falciparum* strain Dd2 from a mixed *Plasmodium* and human sample (30:70 *P.falciparum* Dd2:human). The hybridisation was successful, but the amount of enrichment was lower (67.49% of sequencing reads aligning to the *Plasmodium* genome after enrichment) than was seen when both the probes and DNA library for enrichment were generated from the same strain (Table **3**, **2E**). This was expected because the genomes of the two strains are somewhat different and some regions will not hybridise during the enrichment process. In addition, the enriched DNA was aligned to the 3D7 reference genome and

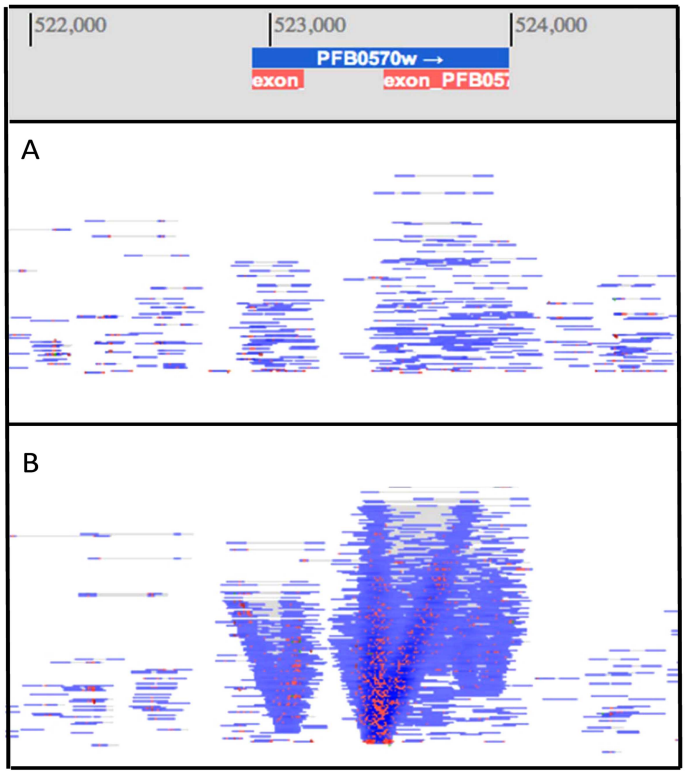

Fig. (3). Plot to compare coverage pre- and post-pulldown for a random exonic, intronic and intergenic region in the *P falciparum* genome [21]. Reads aligned to the region of interest are shown for **A**) a pre-pulldown sample and **B**) a post-pulldown matched sample. In both pre- and post-pulldown samples, coverage of >100 was observed in areas of both of the exonic regions shown. Intergenic and intronic regions show low or no coverage. Blue reads indicate perfect matches to the reference genome and red reads indicate the presence of non-matched bases (less than 2 nucleotides mismatched per read).

Table 4. Coding Region Coverage and SNP Detection for the *P.falciparum* Genome Pre- and Post- solution-based Pulldown

Experiment	Sample ID	Sample	Probe Strain	Average Depth	Maximum Depth	% of Coding Region Covered at 5x	Number of SNPs Detected
2A	2A i	3d7:human 25:75 prepulldown	~	49	548	99.86	9
	2A ii	3d7:human 25:75 800ng	3D7 1.5µg	266	8005	95.04	7
	2A iii	3d7:human 25:75 500ng	3D7 1.5µg	243	8006	73.46	10
2B	2B i	3D7:human 25:75 prepulldown	~	48	479	99.48	13
	2B ii	3D7:human 25:75 1600ng	3D7 1.5µg	145	7948	90.11	12
	2B iii	3D7:human 25:75 1600ng	3D7 3µg	204	7970	93.9	12
	2B iv	3D7:human 25:75 800ng	3D7 1.5µg	317	7987	90.59	13
	2B vi	3D7:human 25:75 800ng	3D7 3µg	224	7971	70.32	12
2C	2C i	3d7:human 30:70 prepulldown	~	32	482	98.85	12
	2C ii	3D7:human 30:70 55°C hyb	3D7 1.5µg	54	3623	40.97	3
	2C iii	3D7:human 30:70 65°C hyb	3D7 1.5µg	63	4625	36.11	10
2D	2D i	3d7:human 20:80 prepulldown	~	36	331	99.33	8
	2D ii	3d7:human 20:80	3D7 1.5µg	311	7989	97.06	10
	2D iii	3d7:human 20:80 1M TMAC	3D7 1.5µg	271	7944	95.65	10
	2D iv	3d7:human 20:80 0.5M TMAC	3D7 1.5µg	317	7981	96.86	10
	2D vi	3d7:human 20:80 0.1M TMAC	3D7 1.5µg	329	7983	98.19	9
2E	2E i	Dd2:human 30:70 prepulldown	~	99	1138	96.44	5599
	2E ii	Dd2:human 30:70	3D7 1.5µg	227	8007	94.58	5158
2F	2F i	3D7:human 20:80 prepulldown	~	19	359	98.92	9
	2F ii	3D7:human 20:80 mix probe	3D7, Dd2, IT and HB3	260	8005	67.15	8
2G	2G i	Isolate 1:human 20:80 prepulldown	~	35	496	96.32	5181
	2G ii	Isolate 2:human 20:80 prepulldown	~	20	259	94.90	4696
	2G iii	Isolate 1:human 20:80 mix probe	3D7, Dd2, IT and HB3	242	8023	79.22	4097
	2G iv	Isolate 2:human 20:80 mix probe	3D7, Dd2, IT and HB3	120	7998	59.39	3310

Average depth refers to the mean number of Illumina GA sequencing reads that can be mapped to each position along the genome and the maximum depth refers to the greatest number of reads mapped to a single location in the genome. Coding region coverage shows the percent of the coding region covered by at least 5 sequencing reads. A minimum SNP quality of 1 error per 1000bp was used for SNP filtering. SNPs were excluded if in telomeric, subtelomeric or non-unique regions or if present in hypervariable gene families. Probe amount refers to the amount of DNA used at the start of probe synthesis.

some sequences specific for the Dd2 strain would not register as *Plasmodium* DNA. This manifests itself as a higher amount of 'other' DNA observed after enrichment for the Dd2 strain. No reference was available for the Dd2 strain. The GC content of the enriched Dd2 DNA was similar to that observed when enriching a *P.falciparum* 3D7 genome (Fig. **2E**). Approximately 95% of the coding region was covered to 5x depth in the enriched sample, and this is similar to the value obtained for the prepulldown sample (96%), indicating that little of the coding region DNA is lost during the hybridisation (Table **4**, **2E**). The average depth of coverage across the exome was approximately double for the postpulldown Dd2 strain sample (227x) compared to the prepulldown sample (99x) and the maximum depth of coverage was approximately 8x higher in the enriched sample (8007x) compared to the non-enriched sample (1138x). Taken together, these results suggest coverage non-uniformity in the enriched sample with some areas having extremely high levels of coverage. SNP calling was carried out for these samples using the *P.falciparum* 3D7 genome as a reference and hence many SNPs were detected both pre- (5599 SNPs) and post-enrichment (5158 SNPs) due to differences between the two strains (Table **4**, **2E**). 92% of the SNPs detected pre-enrichment could also be detected in the enriched sample, showing good SNP detection in these samples.

To further test this technique with other *Plasmodium* strains, probes were made from a mixture of four different *P.falciparum* strains (3D7, Dd2, IT and HB3). This mixed probe was used to enrich for *P.falciparum* reference strain 3D7 DNA from a mixture with human DNA (*Plasmodium*: human 20:80). The four strain probe mix provided as good an enrichment as the single strain probe mix previously used but the coding region showed poorer coverage after enrichment with the mixed probe (Table **3**, **2F** and Table **4**, **2F**). This could possibly suggest that the DNA from the HB3 and IT strains was of poor quality or that these strains are quite divergent in exonic regions and have fewer regions of complimentarity with the 3D7 strain. Despite only 67% of the coding region being covered at 5x depth after enrichment, compared to 99% pre-enrichment, the number of SNPs detected was similar before and after enrichment (Table **4**, **2F**).

This method was also tested using mixtures of human DNA and one of two *P.falciparum* clinical isolates. The two clinical samples used in this study were collected from clinical malaria patients in Southeast Asia and were cultured for a short while before DNA extraction. They are multiclonal in nature and contain a small amount of human contamination even before they were mixed with human DNA (20:80 *P.falciparum* clinical isolate:human) in these experiments. The enrichment of these clinical samples was comparable, albeit not quite as good as for the reference strain, with over 65% of the DNA after pulldown, using the four strain mixed probe, being complimentary to the *P.falciparum* 3D7 reference genome (Table **3**, **2G**). Similar to previous experiments, the AT-rich regions were not captured during the hybridisation and therefore the GC content is not very close to the theoretical value (approximately 35% GC, Fig. **2F**). However, the coding region showed acceptable coverage in both samples post-enrichment (clinical 1 – 79.22% at 5x coverage and clinical 2 – 59.39% at 5x coverage, Table **4**, **2G**), although to a lesser degree than the pre-pulldown samples. Several SNPs were detected in both samples when compared to the 3D7 reference strain (Table **4**, **2G**). SNP detection was poorer after enrichment, with 79% (4097/5181) and 70% (3310/4696) (clinical sample 1 and 2 respectively) of SNPs being detected after pulldown (Table **4**, **2G**). Overall, results from the mixed probe experiments demonstrate that this probe was not very successful for enrichment and this is likely to be due to poor quality of DNA from the IT and HB3 strains. Better results may have been achieved using a pure 3D7 probe with these clinical samples, however this was not attempted. Using a reference strain such as 3D7 to enrich for clinical samples will always miss some regions of the genomes of the clinical isolates due to differences in the genomic sequences. This is a caveat of this kind of experiment and interesting, novel regions from the clinical isolate's genome will be missed, however SNP calling should still be possible.

ALTERNATIVE TECHNIQUE

Due to the enrichment bias observed with the *Plasmodium* genome, an alternative method of enrichment was investigated. In this method, probes were made from whole-genome human DNA and these were used to selectively remove the human DNA from an *E.coli* and human mixed sample. Sequencing was then carried out on the portion of the library that did not hybridise to the probes. It was reasoned that using hybridisation to human DNA could reduce or eliminate the contaminating human DNA whilst maintaining the pathogen genome without bias caused by the effects of different hybridisation efficiencies. However, when using 10µg of human probe, the amount of human sample DNA remaining with the *E.coli* DNA after enrichment was still significant (Table **1**, **3A**). This would suggest that the hybridisation is not efficient and that vast quantities of human probe would be required to successfully remove all traces of human DNA. In addition, the fact that the material used for sequencing is derived from non-hybridised material, means that significant amount of adapter dimer and other contaminants were also sequenced, further reducing sequencing efficiency and *E.coli* enrichment. These poor results did not provide sufficient support to extend this protocol to the *Plasmodium* genome, although further experiments could be carried out.

CONCLUSIONS

In this study a solution-based hybridisation method for the enrichment of pathogen DNA from samples that are heavily contaminated with human DNA was investigated. This technique may have clinical utility in isolating pathogen DNA from clinical samples that are heavily contaminated with human DNA and thus improve whole-genome sequencing efficiency and pathogen identification. The technique developed in this study uses basic molecular laboratory techniques and off-the-shelf reagents. In addition, the probes can be synthesised in-house from whole-genome DNA. This means that the whole process is inexpensive and straightforward, which makes this an attractive alternative to commercial enrichment protocols [2, 3, 9].

The solution-based enrichment technique was successful for the enrichment of *E.coli* DNA from mixed *E.coli*:human samples, even when the initial mix contained as little as 10% *E.coli* DNA. The depth of coverage across the genome was significantly increased by the process, showing that a good level of enrichment was achieved. GC content was not affected by the process and the uniformity of coverage was good throughout the genome. This technique works extremely well for this small, GC-neutral genome.

This technique was also applied to the larger *P.falciparum* genome which is notoriously difficult to sequence compared to many other genomes, largely due to its high AT content (80.6%), particularly in non-coding regions (86.5%) and because of a high content of repetitive sequences [1]. After the enrichment procedure, a significant reduction in the amount of human DNA was observed. However, a bias was observed in the *Plasmodium* DNA that was enriched, with regions with a high AT content being poorly represented after hybridisation. This is likely to be caused by biases within the hybridisation process because AT-rich regions are known to display poorer hybridisation than GC-rich regions [14, 15]. In these experiments, lowering of the hybridisation temperature from 65^{o}C to 55^{o}C or inclusion of the additive TMAC did not reduce this hybridisation bias. The AT bias meant that the uniformity of

coverage across the genome was lower, with many coding regions of the *P.falciparum* genome being extremely over-represented and some intergenic and repetitive regions completely absent. This biased coverage is also seen, although to a lesser extent, with standard next-generation sequencing and therefore recent population genetics studies focus on coding regions. In this study, considering only the coding region, which has a higher GC content than the rest of the genome, improved the coverage, with a large proportion of samples having >90% of their coding regions covered to 5x depth. In addition, SNP calling was similar pre- and -post-enrichment, suggesting a potential clinical utility for this technique. As coverage in coding regions was higher in the post-enrichment samples, SNP detection is therefore of higher confidence.

It is possible that RNA probes similar to those used by Melnikov *et al*., [12] may have different hybridisation kinetics to the DNA probes used here, resulting in better hybridisation to, and capture of, AT- rich sequences. Further investigation is required to determine which of these approaches is superior.

As an alternative approach, probes were made with human DNA and these were used to hybridise and remove human DNA from a mixed sample of human and pathogen DNA. It was found that even when large quantities (10µg) of human DNA was used for probe synthesis, there was still significant amounts of human contamination present with the bacterial (*E.coli*) DNA after the hybridisation-based cleanup. This would suggest that the hybridisation process is very inefficient when using the human probe, probably due to the large size and high complexity of the human genome. It is likely that the amount of human probe required for efficient elimination of human contamination in clinical material may be difficult and expensive to obtain for routine use.

This solution-based hybridisation technique provides a cost-effective, straightforward method to enrich samples for a genome of interest. The method was successfully applied to mixtures of human and *E.coli* DNA, with high amounts of evenly distributed coverage obtained throughout the *E.coli* genome. When applied to the difficult *P.falciparum* genome, an uneven coverage across the genome was obtained, which is likely to be caused by difficulties in hybridisation of the AT-rich regions of the genome. Non-uniformity of coverage was also observed in pre-enrichment samples but to a lesser extent. Despite the lack of uniformity observed in the enriched samples, the coding region of the *P.falciparum* genome showed high levels of coverage and SNP calling was possible and comparable to the non-enriched samples. The utility of this process could be further developed to enable the simultaneous hybridisation of several patient samples through indexing of the patient DNA or it could be used to isolate several different pathogens in the same capture experiment and therefore it may be useful for metagenomic studies.

METHODS

DNA

Male human DNA (Promega) and *E.coli* DNA (strain B, Sigma-Aldrich) were purchased from commercial suppliers. *P. falciparum* DNA was obtained from either i) long-term cultured parasite lines (laboratory clones 3D7, HB3, IT and Dd2) or ii) short-term culture parasite isolates obtained from patients with malaria in Cambodia. *P. falciparum in vitro* culture was performed as previously described [18]. Genomic DNA was extracted from infected red blood cell pellets using the QIAamp DNA Blood MidiKit (Qiagen) according to the manufacturer's protocol.

Clinical samples were collected with written consent from adult patients or from a parent or guardian of paediatric patients. These samples were collected under a project with ethical approval obtained from the Ministry of Health in Cambodia.

PROBE SYNTHESIS

Probes were synthesised from 1.5µg or 3µg of whole-genome *E.coli* or *P. falciparum* DNA samples, or 10µg of whole-genome human DNA. The DNA was fragmented to approximately 150bp by adaptive focused acoustics using a Covaris S2 (Covaris. Inc.) with AFA tubes (Duty cycle: 20, Intensity: 5, Cycle burst: 200, Duration: 180). Subsequently, the DNA fragments were end repaired (1 hour at 20^{o}C) and A-tailed (1 hour at 37^{o}C) following the standard Illumina library preparation protocol and using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs). Intervening purification steps were by column purification (QIAquick or Minelute PCR Purification Kits, Qiagen). The DNA fragments were then enzymatically biotinylated by addition of 0.1mM Biotin-16-dUTP (Biotium), 100U terminal transferase, 1.25µM $CoCl_2$ and 1x terminal transferase buffer (New England Biolabs). Incubation was carried out for 30mins at 37^{o}C prior to column purification using the Minelute PCR Purification kit (Qiagen) and elution in 10µl of EB Buffer (Qiagen).

LIBRARY SYNTHESIS

Libraries were synthesised from a total of 1µg of DNA consisting of a mixture of *E.coli* or *P.falciparum* DNA mixed in differing proportions with human DNA. The DNA was fragmented to approximately 200-300bp using the Covaris S2 with AFA tubes (Duty cycle: 20, Intensity: 5, Cycle burst: 200, Duration: 120). End repair, A-tailing and adapter ligation were carried out following standard Illumina library preparation methodology. The NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs) was used throughout and incubations were carried out for 1 hour at 20^{o}C (end repair), 1 hour at 37^{o}C (A-tailing) and for 2 hours at 18^{o}C (adapter ligation). Standard Illumina adapter sequences (forward 5'-GATCGGAAGAGCGGTTCAGCA GGAATGCCGAG-3', reverse 5'-ACACTCTTTCCCTA CACGACGCTCTTCCGATCT-3') were used. Column purification (QIAquick or Minelute PCR Purification Kits, Qiagen) was carried out between each stage and after the final column purification step the libraries were eluted in 50µl of EB buffer (from the Qiagen kit). PCR amplification was carried out using 2µl Herculase II Phusion DNA Polymerase, 0.1mM dNTP mix and 1x Herculase Reaction Buffer (Stratagene) in a total volume of 100µl. 0.2µM of each of the standard Illumina primer sequences (forward 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCT-3', reverse.

5'-CAAGCAGAAGACGGCATACGAGATCGGTCTC GGCATTCCTGCTGAACCGCTCTTCCGATCT-3') were used per reaction and these were purchased from Sigma-Genosys. Amplification was achieved using the following cycling conditions: 98°C for 30sec followed by six cycles of 98°C for 10sec, 65°C for 30sec, 72°C for 30sec and a final elongation incubation of 72°C for 5mins. After amplification, the libraries were purified using AMPure beads (Agencourt) and eluted in a total of 60µl of water (the beads were eluted twice with 30µl). The Agilent 2100 Bioanalyser (Agilent Technologies) and the associated DNA1000 kit (Agilent Technologies) were used to determine quality and concentration of the libraries.

HYBRIDISATION

500ng, 800ng or 1600ng of each library (determined from the Agilent 2100 Bioanalyser traces) was lyophilised using a Concentrator Plus/Vacufuge Plus (Eppendorf) at room temperature. Once completely dried, the libraries were resuspended in 5µl of ultrapure water. 1µg of each of two blocking oligos (standard Illumina primer sequences, 5'AATGATACGGCGACCACCGAGATCTACACTCTTT CCCTACACGACGCTCTTCCGATCT3'

5'CAAGCAGAAGACGGCATACGAGATCGGTCTCG GCATTCCTGCTGAACCGCTCTTCCGATCT3') was added to the library samples. Library samples were denatured at 96°C for 5 minutes before incubation at 65°C. 40µl aliquots of 2x hybridisation buffer (10x SSPE (Sigma-Aldrich), 10x Denhardt's Solution (Sigma-Aldrich), 10mM EDTA (Calbiochem), 0.2% SDS (Sigma-Aldrich)) underwent a similar denaturation and incubation process. After incubation of the libraries and hybridisation buffer aliquots at 65°C for 5 minutes, one 5µl aliquot of probe solution (half of the original 3µg preparation) for each library was mixed with 5µl of water prior to incubation at 96°C for 5 minutes. The probes and 16µl of hybridisation buffer were then added to the libraries, whilst maintaining the 65°C incubation temperature. Incubation was then carried out for a further 24 hours at this temperature.

ISOLATION OF HYBRIDISED DNA

After incubation was complete, the samples were added to Dynabeads® MyOne™ Streptavidin C1 beads (Invitrogen) that had been pre-washed and resuspended in 200µl of binding buffer (1M NaCl (Sigma-Aldrich), 10mM Tris-HCl pH7.5 (Sigma-Aldrich), 1mM EDTA (Sigma-Aldrich)). The samples were then placed on a Nutating Mixer (VWR International) for 30mins to mix. The beads and solutions were then separated using a magnetic separator and the beads were washed with Wash Buffer 1 (1xSSC (Sigma-Aldrich), 0.1% SDS (Sigma-Aldrich)) before incubation at room temperature for 15 minutes. Wash buffer 1 was then removed and three washes with Wash Buffer 2 (0.1xSSC (Sigma-Aldrich), 0.1% SDS (Sigma-Aldrich)) were carried out. Wash buffer 2 was pre-warmed to 65°C and incubations between each step were carried out for 10 minutes at this temperature. 50µl of 0.1M NaOH (Sigma-Aldrich) was then added to the beads. The beads were mixed thoroughly before magnetic separation and collection of the sample eluate. The eluate was then neutralised by addition of an equal volume of 1M Tris-HCl pH7.5 (Sigma) before the samples were desalted by column purification (QIAquick PCR Purification Kit, Qiagen) and eluted in 36µl of EB buffer. The recipes for many of the buffers used in this protocol were obtained from Gnirke *et al.,* (2009) Nature Biotechnology, 27 [9].

SAMPLE QUANTIFICATION

Samples were quantified both pre- and post-hybridisation by qPCR using the KAPA Illumina ABI Library Quantification Kit (KAPA Biosciences), following the manufacturer's protocol.

SEQUENCING

Pre- and post-hybridisation samples were sequenced using an Illumina GAIIx machine and associated Illumina reagents. Samples were loaded at 12-14pM and a single lane of paired-end 76bp sequencing (except experiment 1A where 37bp paired-end sequencing was used) was carried out for each sample.

DATA ANALYSIS

All read mappings were carried out using the software package BWA and the coverage data were derived from their pileup analysis using the software package SAMtools [19, 20]. Using the same mapping results, GC plots were calculated. The theoretical GC curves were generated from the genome reference using a read generated starting from each base in the reference and of the same length as the sequencing reads.

For *P.falciparum* SNP analysis, sequence reads were aligned to the 3D7 reference genome (version 2.1.5) using BWA. A list of SNPs was generated using SAMtools and BCFtools at default settings. A minimum SNP quality of 30 (1 error per 1000bp) was used to filter the initial list of SNPs. For further analysis, the SNPs that mapped to telomeric, subtelomeric, non-unique regions and highly variable gene families were excluded. Allele calls were compared between samples using the R statistical package. To observe the distribution of coverage and potential location of SNPs on the whole genome, sequencing data was uploaded into LookSeq, a browser-based viewer for deep sequencing data [21].

ABBREVIATIONS

E.coli	=	*Escherichia coli*
Gb	=	Gigabase
NGS	=	Next generation sequencing
P.falciparum	=	*Plasmodium falciparum*
qPCR	=	Quantitative polymerase chain reaction
SNP	=	Single-nucleotide polymorphism
TMAC	=	Tetramethyl ammonium chloride

CONFLICT OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGEMENTS AND FUNDING

The authors would like to thank the Wellcome Trust Sanger Institute Illumina production team for carrying out some of the sequencing runs and all other staff that contributed helpful discussions. The authors would also like to thank Dr. Duong Socheat and Dr. Chea Nguon for supporting the malaria project conducted in Cambodia. This work was supported by the Wellcome Trust (grant number 098051).

SUPPORTIVE/SUPPLEMENTARY MATERIAL

Supporting information is available on the publishers Web site along with the published article.

REFERENCES

[1] Gardner MJ, Hall N, Fung E*, et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 2002; 419: 498-511.

[2] Horz HP, Scheer S, Huenger F, Vianna ME, Conrads G. Selective isolation of bacterial DNA from human clinical specimens. J Microbiol Methods 2008; 72: 98-102.

[3] Horz HP, Scheer S, Vianna ME, Conrads G. New methods for selective isolation of bacterial DNA from human clinical specimens. Anaerobe 2010; 16: 47-53.

[4] Dame JB, McCutchan TF. Plasmodium falciparum: Hoechst dye 33258-CsCl ultracentrifugation for separating parasite and host DNAs. Exp Parasitol 1987; 64: 264-6.

[5] Auburn S, Campino S, Clark TG*, et al.* An Effective method to purify plasmodium falciparum DNA directly from clinical blood samples for whole genome high-throughput sequencing. PLoS One 2011; 6: e22213.

[6] Bernhardt M, Pennell DR, Almer LS, Schell RF. Detection of bacteria in blood by centrifugation and filtration. J Clin Microbiol 1991; 29: 422-5.

[7] Carter V, Cable HC, Underhill BA, Williams J, Hurd H. Isolation of Plasmodium berghei ookinetes in culture using Nycodenz density gradient columns and magnetic isolation. Malar J 2003; 2: 35.

[8] Bainbridge MN, Wang M, Burgess DL*, et al.* Whole exome capture in solution with 3 Gbp of data. Genome Biol 2010; 11: R62.

[9] Gnirke A, Melnikov A, Maguire J*, et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009; 27: 182-9.

[10] Johansson H, Isaksson M, Falk Sorqvist E*, et al.* Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res 2010; 39(2): e8.

[11] Tewhey R, Warner JB, Nakano M*, et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 2009; 27(11): 1025-31.

[12] Melnikov A, Galinsky K, Rogov P*, et al.* Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol 2011; 12: R73.

[13] Blattner FR, Plunkett G 3rd, Bloch CA*, et al.* The complete genome sequence of Escherichia coli K-12. Science 1997; 277: 1453-62.

[14] Marmur J, Doty P. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J Mol Biol 1962; 5: 109-18.

[15] Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res 2006; 34: 564-74.

[16] Chevet E, Lemaitre G, Katinka MD. Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. Nucleic Acids Res 1995; 23: 3343-4.

[17] Melchior WB Jr, Von Hippel PH. Alteration of the relative stability of dA-dT and dG-dC base pairs in DNA. Proceedings of the National Academy of Sciences of the United States of America 1973; 70: 298-302.

[18] Trager W, Jenson JB. Cultivation of malarial parasites. Nature 1978; 273: 621-2.

[19] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25: 1754-60.

[20] Li H, Handsaker B, Wysoker A*, et al.* The Sequence Alignment/ Map format and SAMtools. Bioinformatics 2009; 25: 2078-9.

[21] Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. Genome Res 2009; 19: 2125-32.

Received: January 17, 2011 Revised: March 06, 2012 Accepted: March 07, 2012