NSR-seq transcriptional profiling enables identification of a gene signature of *Plasmodium falciparum* parasites infecting children

Marissa Vignali,1 Christopher D. Armour,2 Jingyang Chen,1,3 Robert Morrison,1 John C. Castle,2 Matthew C. Biery,2 Heather Bouzek,2 Wonjong Moon,1 Tomas Babak,2 Michal Fried,1,4 Christopher K. Raymond,2 and Patrick E. Duffy1,3,4

1Seattle Biomedical Research Institute, Seattle, Washington, USA. 2Department of Molecular Informatics and Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck and Co. Inc., Seattle, Washington, USA. 3Laboratory of Malaria Immunology and Vaccinology, National Institute of Allergy and Infectious Diseases, NIH, Rockville, Maryland, USA. 4Department of Global Health, University of Washington, Harborview Medical Center, Seattle, Washington, USA.

Malaria caused by *Plasmodium falciparum* results in approximately 1 million annual deaths worldwide, with young children and pregnant mothers at highest risk. Disease severity might be related to parasite virulence factors, but expression profiling studies of parasites to test this hypothesis have been hindered by extensive sequence variation in putative virulence genes and a preponderance of host RNA in clinical samples. We report here the application of RNA sequencing to clinical isolates of *P. falciparum*, using not-so-random (NSR) primers to successfully exclude human ribosomal RNA and globin transcripts and enrich for parasite transcripts. Using NSR-seq, we confirmed earlier microarray studies showing upregulation of a distinct subset of genes in parasites infecting pregnant women, including that encoding the well-established pregnancy malaria vaccine candidate var2csa. We also describe a subset of parasite transcripts that distinguished parasites infecting children from those infecting pregnant women and confirmed this observation using quantitative real-time PCR and mass spectrometry proteomic analyses. Based on their putative functional properties, we propose that these proteins could have a role in childhood malaria pathogenesis. Our study provides proof of principle that NSR-seq represents an approach that can be used to study clinical isolates of parasites causing severe malaria syndromes as well other blood-borne pathogens and blood-related diseases.

Introduction

Malaria is endemic in over 100 countries, and 3.3 billion people are at risk of contracting this disease (1). While only a small percentage of clinical cases are fatal, cumulatively they amount to approximately 1 million deaths per year, so that malaria remains one of the leading worldwide causes of mortality. Severe malaria, caused mainly by *Plasmodium falciparum*, is most frequent in young African children who have not yet developed naturally acquired immunity but is also a concern for nonimmune travelers in malaria-endemic areas (2). In addition, adult women living in malarious areas who are otherwise clinically immune become susceptible to a severe form of the disease during pregnancy (3).

Severe malaria might be caused by parasite forms that are more virulent than those that cause mild disease (4). In the case of pregnancy malaria, severe symptoms result from the accumulation of *P. falciparum*-infected red blood cells (pRBC) in the placenta, mediated by binding to host receptor chondroitin sulfate A (CSA) (5). The ability of pRBC to bind to CSA correlates with the export of VAR2CSA/PFL0030c protein to the host cell surface (reviewed in ref. 3). VAR2CSA protein is encoded by 1 of the approximately 60 *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) or var genes present in the parasite genome (6). Other PfEMP1 proteins displayed on the surface of pRBC have been shown to interact with a range of host receptors (such as CD36, ICAM1, VCAM1, PECAM1/CD21, heparan sulfate, etc.), resulting in adhesion to the vascular endothelia (7). Sequestration allows *P. falciparum* to avoid splenic clearance and has been associated with severe syndromes, such as cerebral malaria (7, 8).

PfEMP1 proteins belong to a larger group of variant surface antigens (vsa) that also includes the rifin, stevor, PfMC-2TM, and surfin gene families. Although the functions of these other vsa have not been well defined, they are exported to the host cell surface, and it has been suggested that, like PfEMP1s, they could have a role in protective immunity (9). Hundreds of non-vsa parasite proteins have been shown or predicted to be exported beyond the parasite’s vacular membrane and into the host cell (10, 11). Although little is known about the function of the majority of these exported proteins, some have been implicated in the process of translocation of other parasite proteins, in the creation of parasite-derived structures in the cytoplasm of pRBC, or in the remodeling of the pRBC surface. This extensive remodeling of the host cell has been proposed to be essential for the parasite’s virulence (12).

The identification of genes or proteins related to severe disease could guide intervention strategies, such as vaccines designed to prevent severe malaria and death. For this reason, earlier studies sought to associate the expression of particular parasite genes or proteins with disease severity by comparing the transcriptome and...
proteome of parasites obtained from different types of malaria episodes. Two recent DNA microarray studies failed to identify parasite genes whose expression was related to disease severity (13, 14). In contrast, microarray and proteomic studies of pregnancy malaria have identified sets of genes and proteins, including VAR2CSA, that are more abundant in parasites isolated from pregnant women than in those from children (15, 16) or that are more abundant in placental parasites than in a laboratory-adapted reference strain (17).

In this study, we characterize the transcriptional profile of in vitro cultures of P. falciparum parasites by next-generation sequencing (NGS). Due to its digital nature, NGS is not subject to saturation or cross-hybridization and thus displays a much broader dynamic range than traditional microarrays (18). NGS does not necessarily require a priori knowledge of the underlying genomic sequence and allows the identification of new genes and the refinement of gene models. Therefore, NGS is ideally suited to investigate gene expression in field isolates of P. falciparum, including that of variable surface antigens that are not fully captured by array designs that are necessarily based on previously determined sequences.

The predominant human-derived RNA species in P. falciparum-infected erythrocytes corresponds to globin mRNA (19) and rRNA, which remains a major contaminant even after polyA+ purification (20). To avoid amplification of these very abundant RNAs, which would decrease the complexity of the resulting libraries, we used the not-so-random sequencing (NSR-seq) method (21) to construct NGS libraries from laboratory-adapted strains of P. falciparum cultured in human erythrocytes. We show that this technique excludes the majority of human globin and rRNA and enables the identification of parasite genes whose transcription levels vary among field isolates. In addition, NSR-seq detects most genes previously discovered by microarray analysis as upregulated in parasites infecting pregnant women, including the pregnancy malaria vaccine candidate var2csa (15), demonstrating that this technique is suitable for the identification of differentially regulated parasite genes. More importantly, NSR-seq identifies 4 genes that are expressed preferentially by children’s parasites (glutamatic acid-rich protein [garp/PFA0620c]; mature-parasite-infected erythrocyte surface antigen [MESA/ PfEMP2/PFE0040c]; glycophrin-binding protein 2 [Gbph2/PF13_0010]; and probable protein PF10_0350]), which might have roles of particular importance in the pathogenesis of childhood malaria. All 4 of these genes encode proteins that are known or predicted to be exported beyond the parasitophorous vacuole and into the host cell (22). Exported proteins have been implicated in malaria virulence due to their ability to modify the host cell, including alterations in membrane rigidity; in parasite binding to receptors in the vascular endothelia, the placenta, and other erythrocytes; and in evasion of immune system clearance (reviewed in ref. 23). Garp has been experimentally demonstrated to be very antigenic (24), while the other 3 proteins have high predicted antigenicity (25). Three of the genes upregulated in children’s parasites (garp, Gbph2, and probable protein PF10_0350) are unique to P. falciparum (22), which causes the most severe form of human malaria, suggesting that these proteins could have a role in virulence that is absent in other parasite species. Finally, the expression of these 4 genes has been reported to vary in parasites selected for their ability to bind particular host receptors (26, 27) and in field parasites when compared with laboratory strains (28, 29). Future experiments will attempt to validate these proteins as biomarkers of disease severity and as targets for intervention strategies designed to reduce the severity of childhood malaria. The results described herein lay the foundation for NSR-seq analysis of fresh field isolates of P. falciparum in order to identify other genes whose expression might associate with malaria severity.

**Results**

**NSR-seq of in vitro–cultured P. falciparum reference strain 3D7.** In anticipation of studies with clinical P. falciparum isolates, we modified the recently described NSR-seq protocol (21) and profiled the transcriptome of P. falciparum parasites grown in vitro in human erythrocytes. To deplete highly abundant but uninformative host transcripts, we excluded random hexamer sequences with perfect matches to human 18S and 28S rRNA genes (2,781 hexamers) as well as those that matched human α and β globin mRNA (an additional 384 hexamers). The remaining 931 hexamers (Supplemental Table 1; supplemental material available online with this article; doi:10.1172/JCI43457DS1) were synthesized in the forward and reverse directions of different universal tail sequences to preserve the strand specificity of the sequence tags generated during library construction.

This NSR oligonucleotide set was used as described (21) to generate double-stranded cDNA from polyA+ purified RNA obtained from an unsynchronized in vitro culture of P. falciparum reference strain 3D7 (Supplemental Table 2). The resulting NSR-seq library was submitted to ultra-high-throughput sequencing using a single Illumina (Solexa) GA2 flow cell and generated 36-mer sequence tags or “reads” that were aligned simultaneously to P. falciparum and Homo sapiens sequences using the BowTie algorithm and custom settings that allow, on average, 2 mismatches per read (30). The bioinformatic flowchart used for read assignment is diagrammed in Supplemental Figure 1: after trimming 4 nucleotides off the S′ end (consisting of 2 barcode nucleotides, a random nucleotide, and the first position of the NSR hexamer, which has a higher error rate than the rest of the read), reads that matched an index containing rRNA and globin sequences were removed from further analyses or prefiltered. The remaining reads were first aligned to the parasite and host genomes, and those that did not yield matches were aligned to theoretical transcriptomes of both species in order to capture reads that span splicing sites (see Methods for a detailed explanation). Reads that were not captured in any of these steps were classified as unaligned.

The P. falciparum reference strain 3D7 sample yielded 5.3 million reads, the vast majority of which (97.8%) did not match rRNA or globin genes (Figure 1 and Supplemental Table 3). Over 2.5 million reads (48%) could be assigned unambiguously to a unique location in the parasite or the human genomes and transcriptomes. About 26% of the reads matched multiple parasite and/or human sequences; those that produced 2–30 matches (18.8%) were assigned proportionately to all the corresponding genomic coordinates. Data analysis was performed either exclusively with uniquely aligned reads (“unique reads”) or with reads with 1–30 matches (referred to as “assigned reads”). Assigned reads had an average of 1 mismatch per read: 38.5% of assigned reads aligned to genome sequences with no mismatches, and an additional 27.4% aligned with 1 mismatch (Supplemental Table 3). Finally, reads with more than 30 alignments (7.2%) as well as reads that did not match either parasite or host sequences (23.4%) were excluded from further analyses. The majority of these unaligned reads corresponded to low quality sequences with ambiguous nucleotide calls, while approximately 37% of them matched the...
sequence of oligonucleotides used during library construction and approximately 6% matched non-3D7 \textit{P. falciparum} sequences (Supplemental Table 4). Less than 0.01% of the unaligned reads matched possible contaminants such as \textit{E. coli}, \textit{Mycoplasma spp.}, and homopolymeric sequences.

Overall, the data correlate extremely well with predicted gene models for \textit{P. falciparum}. Over 99% of the reads assigned to the parasite mapped to annotated genes, and only 0.65% mapped to intergenic regions (Supplemental Table 5). The majority of these latter reads extend into the 3' UTRs of predicted genes and are in agreement with EST evidence, while a few are indicative of genes that are not in the current genome annotation or support modification of gene models (Supplemental Figure 2, D–F). The remaining intergenic reads corresponded to singletons and were most likely not indicative of novel genes but resulted from positional biases (see Discussion). The vast majority (96.6%) of reads that align to the parasite genome corresponded to the coding strand of predicted genes, and less than 3.4% mapped to the noncoding strand, confirming that the NSR-seq approach maintains strand-specificity information (Supplemental Table 5). Other important characteristics of this data set, including characteristic gene coverage and detection levels, distribution of reads between coding and noncoding regions, strand specificity, and examples of reads that fall outside of coding regions, are included in Supplemental Figure 2.

**NSR-seq analysis of uninfected human blood.** To estimate the percentage of human sequence reads from the infected blood samples that may be assigned erroneously to parasite genes, we generated and sequenced a NSR-seq transcript library from a sample of polyA+ RNA purified from whole human blood of 5 healthy adult volunteers (Supplemental Table 2). We obtained over 4 million reads, 48% of which could be assigned unambiguously to a single genomic location, while 35% mapped to multiple coordinates and 15% yielded no alignments (Figure 1 and Supplemental Table 3). Almost 3 million reads yielded between 1 and 30 matches and were therefore assigned to coordinates in the parasite and/or human genomes.

Almost 99.9% of the assigned reads matched human sequences exclusively; 0.08% aligned to both human and parasite sequences, and 0.04% aligned exclusively to parasite sequences (Supplemental Table 6). This number is very similar to our estimation of error rate, which revealed that 0.02% of human reads would be assigned to parasite sequences. In comparison, of the 3.5 million assigned reads from the \textit{P. falciparum}-infected erythrocyte sample, 12.4% matched human sequences exclusively, 83.2% matched parasite sequences exclusively, and 4.4% matched both parasite and human sequences. Of the predicted human genes, over 48% received at least 1 read, 32% received at least 10 reads, and approximately 10% received at least 100 reads, whereas less than 7% of the predicted genes in the parasite's genome received at least 1 read, and the number of genes that received 10 or more reads was negligible (Supplemental Table 5). Taken together, these analyses indicate that the erroneous assignment of human reads to parasite sequences and vice versa occurs at a very low frequency.

**Comparison of NSR-seq with other RNA sequencing approaches.** To evaluate the ability of the NSR-seq method to reduce the proportion of human globin and 18S/28S rRNA transcripts from pRBC samples, we compared our data with 2 recently published data sets that also reported RNA sequencing (RNA-seq) of \textit{P. falciparum} parasites cultured in vitro (20, 31). The first study used random nonamers to generate a transcript library from polyA+ RNA (20), while the second study compared different depletion strategies aimed at diminishing the proportion of abundant transcripts from total RNA (31). For each of these studies, we downloaded the reads obtained from the experimental condition that most closely resembled ours (respectively, the \textit{P. falciparum} Not-Marched Control [NMC] sample, ref. 20, and the combined Affinity Depletion/Exonuclease [ADE] sample, ref. 31), and we analyzed them in our bioinformatic pipeline.

The NSR approach generated the highest proportion of aligned reads (74.34%), compared with 52.99% aligned reads for the ADE sample and 17.76% aligned reads for the NMC sample (Supplemental Table 7). Our analysis confirmed that, as originally reported,
over 65% of the reads in the NMC sample could not be aligned to either parasite or human sequences, while the ADE approach yielded only 10.24% unaligned reads, compared with 23.42% unaligned reads in our NSR approach. Next, we counted all reads that matched human or parasite 18S/28S rRNA sequences, independently of whether they were unambiguously assigned to particular chromosomal coordinates or whether they mapped exclusively to either species, since the human and parasite genomes contain multiple copies of the basic rRNA unit, and the sequences of the rRNA genes are highly similar across both species. Although the NMC library was generated from polyA+-purified RNA, over 15% of the reads corresponded to 18S or 28S rRNA genes, whereas only 2.25% of the reads in the NSR library matched 18S or 28S rRNA genes. In spite of the combined affinity/exonuclease depletion approach, the ADE sample contained the largest proportion of reads that matched 18S or 28S rRNA genes (36.75%). This value is higher than the one reported by Otto et al. (31), who only counted reads that aligned to parasite rRNA. Finally, we compared the number of reads that matched human globin in all 3 data sets. Whereas 0.25% of the reads in the NMC sample mapped to globin mRNA, only 0.01% and 0.02% of the reads in the NSR sample and the ADE sample, respectively, corresponded to globin (Supplemental Table 7). These analyses suggest that the NSR-seq protocol efficiently minimizes the number of reads that correspond to globin RNA and rRNA, resulting in the highest proportion of aligned reads.

Next, we investigated whether the use of a reduced set of hexamer sequences influenced our ability to amplify parasite-encoded mRNAs by calculating their ability to bind to all predicted protein-coding *P. falciparum* genes. Over 98% of the protein-coding genes predicted in the parasite’s genome contained more than 20 hexamer-binding sites (HBSs) per 100 bp (Supplemental Figure 3B and Supplemental Table 8), suggesting that the NSR-seq oligonucleotide set is sufficiently complex to amplify the vast majority of *P. falciparum* genes. We also considered excluding an additional 521 hexamers that have perfect matches to *P. falciparum* 18S and 28S. However, we determined that the resulting set of 410 hexamers available for RNA amplification is not sufficiently complex to yield adequate coverage, since, for this further reduced set, none of the parasite genes contained more than 20 binding sites per 100 bp, and over 99% of the genes have fewer than 10 HBSs per 100 bp (Supplemental Figure 3B and Supplemental Table 8).

We also compared relative gene abundance in the NSR, NMC, and ADE data sets, as measured by the normalized number of reads received by each protein-coding gene (Supplemental Figure 3, C and D). The Spearman correlation coefficients between the NSR and NMC data sets and the NSR and ADE data sets, respectively, are 0.63 and 0.66, while the correlation between the NMC and ADE data sets is 0.73. The higher correlation observed between the NMC and ADE data sets stems from the fact that they were obtained from parasites collected 44 and 48 hours after invasion of erythrocytes, respectively (ref. 20 and Llinás et al., personal communication), while the NSR sample corresponds to an unsynchronized culture with a higher proportion of early intraerythrocytic stages. To further illustrate this point, we analyzed the transcriptional profiles of all 3 samples using 2 in-house algorithms (described in Supplemental Methods) that estimated the stage composition of a mixed parasite culture. The first one (Supplemental Figure 3D) estimated the cell cycle progression of parasites measured in hours after invasion of the erythrocyte and is based on published transcriptional expression profiles of in vitro–cultured erythrocytic stages of *P. falciparum* (32). The second one (Supplemental Figure 3E) estimated the contribution of each cell cycle stage to the observed gene expression pattern and is based on a different microarray study that includes nonerythro-
Table 1
List of most differentially regulated genes between cultured maternal and children's field isolates

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Product</th>
<th>No. Reads with 1 match to parasite seq</th>
<th>Abundance in pool Maternal parasites</th>
<th>Abundance in pool Children's parasites</th>
<th>Signal peptide prediction of GPI anchor domains</th>
<th>Export prediction</th>
<th>Antigenicity percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFL0030c</td>
<td>PfEMP1</td>
<td>407</td>
<td>163</td>
<td>1.83</td>
<td>3.67</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PFD1140w</td>
<td>Plasmodium exported protein (PHISTc), unknown function</td>
<td>481</td>
<td>84</td>
<td>3.52</td>
<td>2.52</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>PFB0115w</td>
<td>Conserved Plasmodium protein, unknown function</td>
<td>1,854</td>
<td>523</td>
<td>1.83</td>
<td>1.04</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>MAL7P1.214</td>
<td>PfEMP1 (pseudogene)</td>
<td>87</td>
<td>7</td>
<td>3.56</td>
<td>1.02</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PFL0040c</td>
<td>PfEMP1 (pseudogene)</td>
<td>51</td>
<td>1</td>
<td>4.67</td>
<td>1.87</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PFL0050c</td>
<td>Plasmodium exported protein (PHISTb), unknown function</td>
<td>251</td>
<td>90</td>
<td>1.48</td>
<td>2.04</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PFI0040c</td>
<td>PfEMP1 (pseudogene)</td>
<td>11</td>
<td>3</td>
<td>1.81</td>
<td>1.02</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PFL0940c</td>
<td>PfEMP1 (pseudogene)</td>
<td>20</td>
<td>4</td>
<td>2.20</td>
<td>1.87</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PF13_0045</td>
<td>S-antigen</td>
<td>25</td>
<td>4</td>
<td>2.75</td>
<td>2.04</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>MAL13P1.470</td>
<td>Plasmodium exported protein (PHISTa), unknown function</td>
<td>58</td>
<td>22</td>
<td>1.39</td>
<td>1.04</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PFA0620c</td>
<td>Garp</td>
<td>378</td>
<td>2,038</td>
<td>–2.43</td>
<td>5.07</td>
<td>&lt;3</td>
<td>0</td>
</tr>
<tr>
<td>PF10_0343</td>
<td>MSP3</td>
<td>275</td>
<td>893</td>
<td>–1.77</td>
<td>4.55</td>
<td>&lt;3</td>
<td>0</td>
</tr>
</tbody>
</table>

For each gene, the table includes unique identifiers (Gene ID) and a brief description of the gene product. Genes that were identified as upregulated in maternal parasites by microarray are indicated (15). The table also includes the normalized abundance in the maternal and children's parasite pools in RPKM units (46), using both reads with a unique match and reads with 30 or fewer matches to genomic sequences and the corresponding fold changes (log$_2$ maternal/children's) and $P$ values calculated as described previously (21). Predictions of signal peptides, transmembrane domains, and export to the host were obtained using Signal P (where a score > 3 is considered significant), TMHMM2, and ExportPred (score ≥ 5), respectively (22). Predictions of GPI anchors and antigenicity percentile (indica
The analyses confirmed that the NMC and ADE samples are enriched for late intraerythrocytic cycle (IEC) stages, while the NSR sample corresponds to a mixed population containing a higher proportion of early IEC stages.

Finally, we estimated gene coverage by comparing the number of bases in exons that received reads in all 3 data sets (Supplemental Table 9). While the NSR data set had the largest number of aligned reads, the ADE approach resulted in more exonic bases receiving reads (e.g., 52.1% of bases in exons received at least 1 read for ADE, vs. 13.5% for NSR and 12.6% for NMC; 16.7% received at least 5 reads for ADE, vs. 5% for NSR and 2% for NMC). This is reflected in the less uniform coverage yielded by NSR-seq, although this approach maintains strand specificity, in contrast to the 2 other standard RNA-seq approaches that have been applied to *P. falciparum* (Supplemental Figure 4A). Importantly, the NSR-seq read alignment pattern is highly reproducible and independent of expression level (ref. 21 and Supplemental Figure 4B), thus allowing comparison of expression of the same transcript between different samples.

These results demonstrate that (a) the NSR-seq protocol greatly reduces the proportion of the total available sequence space consumed by uninformative human rRNA and globin mRNA reads, thereby maximizing the proportion of reads that correspond to protein-coding parasite genes of interest; (b) NSR-seq easily differentiates parasite and human transcripts; and (c) the use of NSR oligonucleotides for construction of the transcript library does not limit our ability to measure the expression of the vast majority of parasite genes, although coverage along individual genes is not as uniform as with standard RNA-seq protocols. Taken together, these observations suggest that NSR-seq constitutes an effective tool for comparing the expression of *P. falciparum* genes between samples in the presence of human transcripts.

Based on these results, we next applied NSR-seq to identify differentially regulated genes among clinical isolates. We chose to compare the transcriptomes of parasites that had been isolated from the peripheral blood of pregnant women and children and recently adapted to short-term in vitro culture. Previous microarray analyses of comparable samples identified a small subset of genes upregulated in maternal parasites (15) and provided a basis for comparison with NSR-seq analyses. In order to obtain sufficient quantities of polyA+ RNA for library construction, we pooled RNA purified from in vitro culture-adapted field isolates recently collected from 6 pregnant women and 8 children, respectively (Supplemental Table 2). Thin blood smears revealed that 56% of the pRBC in the pooled maternal parasite cultures and 59% of those in the pooled children’s parasite cultures contained ring-stage parasites; that 41% and 37% of the pRBC, respectively, contained trophozoites; and that only 3% of the pRBC in either pool contained schizonts. In addition to this microscopic evaluation, comparison of the gene expression profiles to available microarray data (32, 33)
from in vitro–cultured *P. falciparum* strains using the algorithms described above suggested that these 2 samples had almost identical contributions from different IEC stages of the parasite’s life cycle (Supplemental Figure S5, A and B).

The maternal parasite sample yielded 10.3 million reads, while the sample corresponding to children’s parasites yielded 3.3 million reads (Figure 1 and Supplemental Table 3). Over 86% of these reads had no or only 1 mismatch to human or parasite sequences. Between 58%–60% of the reads aligned unambiguously to 1 location, 10%–11% of the reads aligned to multiple locations, and 20%–22% of the reads did not align to any location. The lower number of reads assigned to multiple locations as compared with that of the unsynchronized *P. falciparum* reference strain 3D7 sample correlates with the very low amount of human material in the field isolate samples (less than 1%; Supplemental Table 6). Approximately 8% of the reads aligned to 18S or 28S rRNA, and less than 0.01% of the reads aligned to globin.

Based on the number of aligned parasite reads, most genes were not differentially regulated between maternal and children’s isolates (Figure 2, Supplemental Figure SC, and Supplemental Table 10). A small subset of genes departs from the line of equivalence, indicating different levels of transcript in the sample pools. To describe the patterns related to the presence of garp in children’s parasites, we compared the 2 samples to confirm the 6 most highly upregulated genes of children’s parasites (Supplemental Table 13), suggesting that NSR-seq constitutes a more sensitive approach for the identification of differentially regulated genes in these samples.

Expression of *garp* in children using microarray analyses (15). The remaining 3 genes encode S-antigen (PF10_0343), whose sequence is known to vary widely between isolates (36), putative 40S ribosomal protein S27 (PF13_0045), and a protein of unknown function (PFE0495w). The 8 genes that were most significantly upregulated in children’s parasites include *garp*; *Gbph2*; *MESA*; *PF10_0350*, which encodes a probeable protein of unknown function; *cytoadherence-linked asexual gene 3.1* (*Clag3.1/PFC0120w*); merozoite surface protein 3 (*MSP3/PF10_0345*); and *PFA0700c* and *PFD0080c*, which encode 2 exported proteins of unknown function that belong to the hyp10 and hPHTb protein families, respectively. Reanalysis of our published microarray study that compared parasite gene expression from peripheral blood samples obtained from malaria-infected pregnant women and children (15) showed that only *garp* and probable protein *PF10_0350* are significantly upregulated in children’s parasites (Supplemental Table 13), suggesting that NSR-seq constitutes a more sensitive approach for the identification of differentially regulated genes in these samples.

**Discussion**

Malaria disease might be influenced by the ability of parasites to express specific subsets of proteins, as has been demonstrated for the parasites causing pregnancy malaria (3). Therefore, the identification of genes or proteins whose increased expression is indicative of disease severity could guide intervention strategies aimed at lessening the symptoms of malaria in nonimmune and partially immune individuals. Microarray-based transcriptional analyses have been productive in the characterization of proteins that associate with pregnancy malaria parasites (15) but to date have failed to identify parasite genes related to malaria disease in children (13, 14).

In this study, we successfully adapted the recently developed NSR-seq method (21) to profile *P. falciparum* gene expression in human blood samples. To evaluate the ability of NSR-seq to identify differentially regulated genes, we compared the transcriptomes of parasites collected from pregnant women versus children, soon after being adapted to in vitro culture. Comparable isolates have been previously analyzed by microarray assays by our group, yielding a subset of genes whose increased expression associates with maternal parasites (15). The robustness of the NSR-seq technique is demonstrated by its ability to identify the majority of these maternal parasite genes. The list of the genes that showed the highest upregulation in maternal parasites includes *var2csa* (7.6 fold, *P = 9.04 × 10^{-4})*, which encodes the best-characterized pregnancy malaria vsa (17, 34, 35), as well as 4 other PFEMP1 family members that are pseudogenes in the reference strain 3D7. Fur-
Table 2 | Characteristics of genes upregulated in children’s parasites

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Product</th>
<th>Evidence of antigenicity</th>
<th>Evidence of export</th>
<th>Evidence of membrane localization</th>
<th>Functional annotation</th>
<th>Expression data</th>
<th>Conservation</th>
<th>Other functional data</th>
</tr>
</thead>
</table>
| PFA0620c | gapr    | Very high experimental (24) and predicted (25) | Predicted (FEXEL) | Predicted (TMD) | Possible function within the infected erythrocyte | Higher expression in CD36- vs. CSA-binding parasites (27); higher expression in field vs. laboratory-adapted parasites (26); higher expression in CD36-binding parasites vs. lab-adapted parasites (26) | Unique to Plasmodia species (22), suggesting that its role is only required in P. falciparum, which is the only human malaria parasite capable of sequestering and causing severe symptoms. The expression of gapr has also been reported to be higher in parasite isolates that bind host receptor CD36 when compared with parasites that bind CSA (27). The majority of parasites isolated from children bind CD36, while placental parasites always bind CSA (5). Finally, gapr is expressed at higher levels in field parasites than in P. falciparum reference strains that have been adapted to long-term in vitro culture, suggesting that the function of this protein is not required outside of the human host (28). These data taken together suggest that gapr protein carries out a function in the pRBC that is most important in childhood malaria and might be related to the ability of these parasites to bind to specific host receptors. Since gapr protein is not predicted to possess a transmembrane domain or a glycosylphosphatidylinositol (GPI) anchor (although it is predicted to have a signal peptide) (22), further experiments are required to address its localization in the host cell as well as to elucidate the mechanism by which it might be involved in differential adhesion to receptors and pathogenesis. The second validated gene whose expression is increased in children’s parasites corresponds to MESA. MESA encodes an exported protein that is predicted to have moderately high antigenicity (25) and is known to localize to the internal side of the erythrocyte surface, at which it interacts with the host cell membrane cytoskeleton protein band 4.1 (38, 39). The prevention of this interaction results in the cytoplasmic accumulation of MESA and leads to parasite death (39). The MESA gene was not found to be differ-
entially expressed among children’s and maternal samples in our reanalysis of existing microarray data (Supplemental Table 13). However, MESA peptides were detected using LC-MS/MS at a significantly higher level in parasites collected from children versus pregnant women (Supplemental Table 13). MESA gene abundance was much higher in CD36-binding versus CSA-binding parasites and was inversely related to that of var2cesa (27). Although MESA protein levels reflected this difference, the subcellular localization of this protein was not altered in the parasites with different binding profiles. MESA might participate in the pathogenesis of childhood malaria by modifying parasite adhesion via its interaction with components of the host cytoskeleton that lie beneath the surface of the pRBC.

Another gene whose increased expression characterizes children’s parasites is Gbph2 (40). The protein encoded by this gene belongs to the *P. falciparum*-specific family of glycophorin-binding proteins, together with glycophorin-binding protein 130 precursor (GBP130/PF10_0159) and glycophorin-binding protein homologue (Gbph/PF14_0010) (22). All 3 family members are characterized by the presence of multiple tandem repeats of a 50–amino acid motif in the carboxy-terminal region, which, in GBP130, has been shown to be responsible for the interaction with host glycophorin during invasion of the host cell by merozoites (41, 42). Gbph2 is predicted to be exported beyond the confines of the parasitophorous vacuole (22) and to be among the top 20% of parasite proteins sorted by their predicted antigenicity (25), but its function (as well as that of Gbph) remains uncharacterized. Our previous microarray data (15) did not detect a statistically significant difference in the expression of Gbph2 gene in maternal and children’s parasites (Supplemental Table 13). Peptides corresponding to this protein were not detected in our proteomic studies, so its function (as well as that of Gbph) remains uncharacterized. Our previous microarray data (15) did not detect a statistically significant difference in the expression of Gbph2 gene in maternal and children’s parasites (Supplemental Table 13). Peptides corresponding to this protein were not detected in our proteomic studies, so we cannot assess whether Gbph2 protein is more abundant in children’s parasites. Gbph2 expression has been reported to be more than 10-fold higher in parasites selected for CD36 binding when compared with those selected for their ability to undergo rosetting (binding of pRBC to uninfected erythrocytes) (26).

Finally, we also determined that probable protein PF10_0350 is enriched in children’s parasites. This gene encodes a protein of uncharacterized function that is predicted to have a signal peptide, a GPI anchor, and high antigenicity (22, 25). We did not detect peptides corresponding to this protein using mass spectrometry, but reanalysis of our existing microarray data sets (15) indicated that probable protein PF10_0350 expression levels are significantly higher in parasites isolated from children’s versus those isolated from pregnant mothers (log_2 fold = −3.15, P = 4.66 × 10^-6; Supplemental Table 13). Probable protein PF10_0350 was previously reported to be downregulated in placental parasites when compared with parasites from nonpregnant women and children (17) and has been reported to be expressed at higher levels in field isolates than in laboratory-adapted strains (29), suggesting that its function might be important for binding of pRBC to nonplacental receptors in the human host. As with garp and Gbph2, this protein has no orthologs in other *Plasmodium* species (although it does have a paralog in *P. falciparum*, predicted exported protein of unknown function MAL7P1.171) (22). Little else is known about this gene, although it was recently reported that its mutation by transposon-based insertion resulted in a moderate attenuated growth of the corresponding parasite line, suggesting that its function is important for parasite survival (43). Further experiments, including the experimental verification of the predicted export of probable protein PF10_0350 to the host cell and analysis of its subcellular localization, are needed to better understand the potential role of this protein in childhood malaria pathogenesis.

As in all cDNA library-based approaches, a major concern when constructing RNA-seq libraries is the overrepresentation of certain transcripts, which can severely affect the coverage of genes of interest. Our approach to expression profiling is designed to exclude the majority of human globin and rRNA, which constitute the predominant RNA species in the sample, allowing the analysis of parasite gene expression in the presence of human cells. We first validated the technique on reference strain 3D7: more than two-thirds of the 5 million reads obtained could be assigned to the parasite or human genomes and almost 84% matched parasite sequences exclusively. Furthermore, we demonstrate that the assignment of parasite reads to human sequences and vice versa occurs at a very low frequency. In comparison with earlier studies using random oligonucleotides to analyze the gene expression of *P. falciparum* reference strain 3D7 by RNA-Seq (20, 31), the use of NSR oligonucleotides resulted in the largest proportion of aligned reads and the lowest percentages of reads matching globin and rRNA, while allowing the detection of the vast majority of *P. falciparum* genes (Supplemental Figure 3 and Supplemental Tables 7 and 8). These results indicate that NSR-seq is an effective tool for profiling parasite gene expression in the presence of human transcripts that should be applicable to the profiling of fresh parasite samples.

In conclusion, NSR-seq represents a robust approach for whole-genome expression profiling of *P. falciparum* parasites that overcomes the limitations of earlier methods. In studies of the 3D7 reference *P. falciparum* clone and of recently adapted field isolates, NSR-seq yielded highly specific mRNA transcriptional information, as evidenced by overwhelming alignment of sequences to
annotated genes in the correct orientation and limited assignment to intergenic, rRNA, and human globin sequences. Our sequencing and analysis pipeline correctly identified genes that are known to be preferentially expressed by pregnancy parasites, including a variant antigen gene, and revealed for genes expressed preferentially by children’s parasites. Notably, the genes that distinguish children’s parasites encode proteins of high antigenicity with putative export sequences that would direct them to the host-cell interface, and their variable transcription profiles in strains that differ in their binding properties suggest they could be involved in adhesion and sequestration. Further experiments will be required to better understand the potential roles of these proteins in malaria pathogenesis and immunity. Proteins that associate with severe disease could be targets for vaccines that prevent severe malaria or death. The new NSR-seq technology will be valuable for understanding whether differences in parasite gene expression may explain, in part, the differences in children’s outcomes, similar to the distinct transcriptome of *P. falciparum* parasites that cause placental malaria and disease during pregnancy.

**Methods**

**Sample origin.** Parasite samples were donated by Tanzanian women (aged 16–45 years) and/or their children presenting at the Muheza Designated District Hospital or the Morogoro Regional Hospital in Tanzania. Sample donors were among those enrolled in ongoing cohort studies known locally as the Mother-Offspring Malaria Studies (MOMS). Women provided signed informed consent for themselves and/or their offspring before joining the study. Ethical approval of human care and use of samples for these experiments was obtained from the Western Institutional Review Board in Olympia, Washington, USA, and the National Medical Research Coordinating Committee in Dar es Salaam, Tanzania.

Fresh human blood used in these studies was provided by employee volunteers from Seattle BioMed after they provided written informed consent, under a protocol known as “SBBI Blood Donor Program,” approved by the Western Institutional Review Board.

**Parasite culture and RNA isolation.** *P. falciparum* reference strain 3D7 as well as field isolates recently collected from malaria-infected women and children were cultured in vitro in erythrocytes isolated from whole blood as previously described (44). Red blood cells were collected by centrifugation, and the pellet was mixed with 5 volumes of TRIzol (Invitrogen) and stored at −80°C until RNA purification. The same procedure was applied to blood collected from uninfected adult volunteers.

Samples were thawed at room temperature and total RNA was isolated using the RNeasy Micro Kit (Qiagen) according to the manufacturer’s instructions. Total RNA from uninfected samples as well as samples obtained from the peripheral blood of children and pregnant women infected with *P. falciparum* were pooled as described in Supplemental Table 2 and then subjected to polyA+ RNA purification using the PolyATract Kit (Promega). RNA concentration and quality was determined using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies Inc.) and a 2100 Bioanalyzer (Agilent Technologies Inc.), respectively.

**Oligonucleotides.** NSR primers were designed to deplete human cytoplasmic 18S and 28S rRNA and human globin mRNA sequences during cdNA library construction. We computationally identified 931 distinct hexamers without perfect complementarity to any of the following sequences: human 18S rRNA (U13369.1, nt 3657–5527), 28S rRNA (U13369.1, nt 7935–12969), ε globin mRNA (NM_000558.3, NM_000517.4), and β globin mRNA (NM_000518.4), as obtained from NCBI. Each NSR hexamer was synthesized individually in the antisense and sense orientation with universal 5′ tail sequences to facilitate first-strand (5′-TCCGATCTCTN-[NSR reverse complement]-3′) and second-strand (5′-TCCGATCTGAN-[NSR]-3′) synthesis, respectively. Desalted NSR primers were adjusted to 100 μM with water before pooling. Terminal sites for sequencing by synthesis were added during PCR using distinct forward (5′-AATGATACGGCCGACCGGA-CACCTCTTCCCTTCACGGACGCJTCTCGATCTCT-3′) and reverse (5′-CAACGCAGAGCCACGCAGAGCTCTTCCGATCTGTA-3′) primers. All oligonucleotides were synthesized by Operon Biotechnologies Inc.

**Library generation and sequencing.** Transriptome libraries were constructed with NSR primers, as previously described, with minor modifications (21). First-strand synthesis was performed on 150 μg polyA+ purified mRNA using 10 μM NSR primers and 200 units of Superscript III Reverse Transcriptase (Invitrogen) at 45°C for 90 minutes. The RNA template was subsequently degraded by RNase H (Invitrogen), and cDNA was purified using the QIAquick PCR Purification Kit (Qiagen). All other steps, including second-strand synthesis and PCR amplification, were carried out as described previously. Directional, single-end sequencing of PCR products was performed using the Illumina GAII platform to generate 36-nt reads.

**Bioinformatic analyses.** Reads were assigned to human and parasite sequences using the R software package “DuffyRNAsq,” which is used around the Bowtie algorithm (30) and can be downloaded from http://duffyrnasq.sourceforge.net/. The method is described in detail in the Supplemental Methods and diagrammed in Supplemental Figure 1. Algorithms used to estimate cell cycle stage composition were based on published transcriptional expression profiles of stages of *P. falciparum* (32, 33) and are described in the Supplemental Methods.

**Analysis of existing microarray data sets.** Intensity values for all probes were extracted separately from the Cy3 and Cy5 channels for slides corresponding to all the mother/child pairs included in reference 15, using the R LIMMA package and applying the normalized exponential background correction option. After RNA quantile normalization, all probes for each gene were compared for all children samples versus all maternal samples using a 2-sample t test to derive mean intensities, fold changes, and P values.

**qRT-PCR verification of candidates.** Total RNA extracted from field isolates was amplified with the Ambion MessageAmp II Kit, and qRT-PCR was performed as described previously (15). The oligonucleotides used for amplification are listed in Supplemental Table 11 and were designed using NCBI’s Primer BLAST software or obtained from the literature (34). Relative quantification of RNA expression was calculated by the comparative Ct method of subtracting the Ct of the gene of interest from that of *P. falciparum* seryl-trRNA synthetase (PF07_0073) as a housekeeping gene control after normalizing the amount of cDNA in each qPCR reaction, as described previously (45). An RNA sample prepared from a laboratory-adapted *P. falciparum* isolate was included in all plates as a calibrator. For samples in which expression of a particular gene was undetectable, the Ct value was manually set to 40.

**Mass spectrometry analyses.** Parasite samples collected from malaria-infected women and children were enriched for membrane fractions and processed by tandem mass spectrometry as previously described (16).

**Statistics.** For each protein of interest, a contingency table was generated indicating whether peptides were detected or not in each parasite sample obtained from children or pregnant women. Pearson’s χ² test was used to determine significance.

**Acknowledgments**

Theonest Mutabingwa led the team of MOMS Project clinicians and nurses who collected the clinical samples used in this study. We thank Fenella Raymond for suggesting the collaboration that led to this work; Christopher Mendez, Kathryn Williamson, and Emily Amos for culturing parasites; and Jason Wendler, Cate Speake, and Andrew Oleinikov for helpful discussions. This work was funded in part by a grant from the Foundation for the NIH through the Grand
Challenges in Global Health initiative, the Bill and Melinda Gates Foundation (grant 29202), and the National Institute of Allergy and Infectious Diseases (grant R01AI052059 to P.E. Duffy).

Received for publication April 23, 2010, and accepted in revised form December 15, 2010.

Address correspondence to: Patrick E. Duffy, Laboratory of Malariology and Vaccinology, National Institute of Allergy and Infectious Diseases, NIH, 5640 Fishers Lane, Rockville, Maryland 20852, USA. Phone: 301.443.4605; Fax: 301.480.1959; E-mail: p.earl.duffy@nih.gov.


17. Christopher D. Armour, Matthew C. Biery, and Christopher K. Raymond’s present address is: NuGEN Technologies Inc., Research and Development, San Carlos, California, USA.

18. John C. Castle’s present address is: Institute for Translational Oncology and Immunology, Mainz, Germany.

19. Heather Bouzek’s present address is: Department of Microbiology, University of Washington, Seattle, Washington, USA.

20. Tomas Babak’s present address is: Merck Research Labs, Boston, Massachusetts, USA.


