A novel computational pipeline for var gene expression augments the discovery of changes in the Plasmodium falciparum transcriptome during transition from in vivo to short-term in vitro culture

Clare Andradi-Brown1,2,3, Jan Stephan Wichers-Misterek4-6, Heidrun von Thien4-6, Yannick D. Höppner4-6, Judith A. M. Scholz4, Helle Smedegaard Hansson7, Emma Filtenborg Hocke7, Tim-Wolf Gilberger4-6, Michael F. Duffy8, Thomas Lavstsen7, Jake Baum2,9, Thomas D. Otto10*, Aubrey J. Cunnington1,3*, Anna Bachmann4,6,11,12*, Helle Smedegaard Hansson7, Emma Filtenborg Hocke7, Tim-Wolf Gilberger4-6, Michael F. Duffy8.

1 Section of Paediatric Infectious Disease, Department of Infectious Disease, Imperial College London, UK
2 Department of Life Sciences, Imperial College London, South Kensington, London, SW7 2AZ, UK
3 Centre for Paediatrics and Child Health, Imperial College London, UK
4 Bernhard Nocht Institute for Tropical Medicine, Bernhard-Nocht-Strasse 74, 20359 Hamburg, Germany;
5 Centre for Structural Systems Biology, Hamburg, Germany, Notkestraße 85, 22607 Hamburg, Germany;
6 Biology Department, University of Hamburg, Hamburg, Germany;
7 CMP, University of Copenhagen, 2200 Copenhagen, Denmark;
8 Department of Microbiology and Immunology, University of Melbourne, Melbourne/Parkville VIC 3052, Australia;
9 School of Biomedical Sciences, Faculty of Medicine & Health, UNSW, Kensington, Sydney, 2052, Australia
10 School of Infection & Immunity, MVLS, University of Glasgow, UK
11 German Center for Infection Research (DZIF), partner site Hamburg-Borstel-Lübeck-Riems, Germany

*Equal contribution
¥To whom correspondence should be addressed

Corresponding authors:
Aubrey Cunnington, Section of Paediatric Infectious Disease, Department of Infectious Disease, Imperial College London, Norfolk Place, W2 1PG, London, UK. a.cunnington@imperial.ac.uk
Anna Bachmann, Bernhard Nocht Institute for Tropical Medicine, Bernhard-Nocht-Strasse 74, 20359 Hamburg, Germany. bachmann@bni-hamburg.de
Abstract
The pathogenesis of severe *Plasmodium falciparum* malaria involves cytoadhesive microvascular sequestration of infected erythrocytes, mediated by *P. falciparum* erythrocyte membrane protein 1 (PfEMP1). PfEMP1 variants are encoded by the highly polymorphic family of var genes, the sequences of which are largely unknown in clinical samples. Previously, we published new approaches for var gene profiling and classification of predicted binding phenotypes in clinical *P. falciparum* isolates (Wichers *et al.*, 2021), which represented a major technical advance. Building on this, we report here a novel method for var gene assembly and multidimensional quantification from RNA-sequencing that even outperforms the earlier approach of Wichers *et al.*, 2021 on both laboratory and clinical isolates across a combination of metrics. It is a powerful tool to interrogate the var transcriptome in context with the rest of the transcriptome and can be applied to enhance our understanding of the role of var genes in malaria pathogenesis. We applied this new method to investigate changes in var gene expression through early transition to *in vitro* culture, using paired sets of *ex vivo* samples from our previous study, cultured for up to three generations. In parallel, changes in non-polymorphic core gene expression were investigated. Unpredictable var gene switching and convergence towards var2csa were observed in culture, along with differential expression of 19% of the core transcriptome between paired *ex vivo* and generation 1 samples. Our results cast doubt on the validity of the common practice of using short-term cultured parasites to make inferences about *in vivo* phenotype and behaviour.
Introduction
Malaria is a parasitic life-threatening disease caused by species of the *Plasmodium* genus. In 2021, there were an estimated 619,000 deaths due to malaria, with children under 5 accounting for 77% of these (WHO, 2022). *Plasmodium falciparum* causes the greatest disease burden and most severe outcomes. *P. falciparum* has several highly polymorphic variant surface antigens (VSA) encoded by multi-gene families, with the best studied high molecular weight *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) family of proteins known to play a major role in the pathogenesis of malaria (Leech et al., 1984, Wahlgren et al., 2017). About 60 var genes per parasite genome encode different PfEMP1 variants, which are exported to the surface of parasite-infected erythrocytes, where they mediate cytoadherence to host endothelial cells (Leech et al., 1984, Su et al., 1995, Smith et al., 1995, Baruch et al., 1995, Rask et al., 2010). Var genes are expressed in a mutually exclusive pattern, resulting in each parasite expressing only one var gene, and therefore one PfEMP1 protein, at a time (Scherf et al., 1998). Due to the exposure of PfEMP1 proteins to the host immune system, switching expression between the approximately 60 var genes in the genome is an effective immune evasion strategy, which can result in selection and dominance of parasites expressing particular var genes within each host (Smith et al., 1995).

The polymorphic var genes were classified into four categories (A, B, C, and E) according to their chromosomal location, transcriptional direction, type of 5′-upstream sequence (UPSA–E) (Figure 1a and b) (Lavstsen et al., 2003, Kraemer & Smith, 2003, Kyes et al., 2007, Rask et al., 2010). PfEMP1 proteins have up to 10 extracellular domains, with the N-terminal domains forming a semi-conserved head structure complex typically containing the N-terminal segment (NTS), a Duffy binding-like domain of class α (DBLα) coupled to a cysteine-rich interdomain region (CIDR). C-terminally to this head structure, PfEMP1 exhibit a varying but semi-ordered composition of additional DBL and CIDR domains of different subtypes (Figure 1c). The PfEMP1 family divide into three main groups based on the receptor specificity of the N-terminal CIDR domain. PfEMP1 with CIDRα1 domains bind EPCR, while PfEMP1 with CIDRα2-6 domains bind CD36 and the atypical VAR2CSA PfEMP1 bind placental CSA (Salanti et al., 2004). In addition to these, a subset of PfEMP1 have N-terminal CIDRβ/γ/δ domains of unknown function. This functional diversification correlates with the genetic organization of the var genes. Thus, UPSA var genes encode PfEMP1 with domain sub-variants NTSA-DBLα1-CIDRα1/β/γ/δ, whereas UPSB and UPSC var genes encode PfEMP1 with NTSB-DBLα0-CIDRα2-6. One exception to this rule is the B/A chimeric var genes, which encode NTSB-DBLα2-CIDRα1 domains. The different receptor binding specificities are associated with different clinical outcome of infection. Pregnancy-associated malaria is linked to parasites expressing VAR2CSA, whereas as parasites expressing EPCR-binding PfEMP1 are linked to severe malaria and parasites expressing CD36 binding PfEMP1 are linked to uncomplicated malaria (Turner et al., 2013, Lavstsen et al., 2012, Avril et al., 2012, Claessens et al., 2012, Tonkin-Hill et al., 2018, Wichers et al., 2021). The clinical relevance of PfEMP1 with unknown head structure functions and the PfEMP1 C-terminal domains is largely unknown, albeit both specific endothelial receptor and plasma protein interactions have been described (Tuikue Ndam et al., 2017, Quintana et al., 2019, Stevenson et al., 2015). Each parasite genome carries a similar repertoire of var genes, which in addition to the described variants include the highly conserved var1 pseudogene, which in most genomes occur with no or a truncated exon 2. Also, most genomes carry the unusually small and highly conserved var3 genes, of unknown function.
Comprehensive characterisation and quantification of var gene expression in field samples have been complicated by biological and technical challenges. The extreme polymorphism of var genes precludes a reference var sequence, some var genes are lowly expressed or not expressed at all, contain repetitive domains and can have large duplications (Otto et al., 2019). Consequently, most studies relating var gene expression to severe malaria have relied on primers with restricted coverage of the var family, use of laboratory-adapted parasite strains or have predicted the downstream sequence from DBLα domains (Sahu et al., 2021, Storm et al., 2019, Shabani et al., 2017, Mkumbaye et al., 2017, Kessler et al., 2017, Bernabeu et al., 2016, Jespersen et al., 2016, Lavstsen et al., 2012). This has resulted in incomplete var...
gene expression quantification and the inability to elucidate specific or detect atypical var sequences. RNA-sequencing has the potential to overcome these limitations and secure a better link between var expression and PfEMP1 phenotype in in vitro assays, co-expression with other polymorphic gene families, core genes and epigenetics, which the qPCR and DBLα-tag approaches are not suited for. More recent approaches for var assembly and quantification, based on RNA-sequencing, represent huge advancements compared to qPCR and DBLα-tag analyses. However, they still produce inadequate assembly of the biologically important N-terminal domain region, have a relatively high number of misassemblies and do not provide an adequate solution for handling the conserved var variants (Tonkin-Hill et al., 2018, Wichers et al., 2021, Guillochon et al., 2022).

Plasmodium parasites from human blood samples are often adapted to or expanded through in vitro culture to provide sufficient parasites for subsequent investigation of parasite biology and phenotype (Brown & Guler, 2020). This is also the case for several studies assessing PfEMP1 phenotype of parasites isolated from malaria-infected donors (Pickford et al., 2021, Joste et al., 2020, Storm et al., 2019, Tuikue Ndam et al., 2017, Bruske et al., 2016, Claessens et al., 2012, Lavstsen et al., 2005, Jensen et al., 2004, Kirchgatter & Portillo Hdel, 2002, Dimonte et al., 2016, Hoo et al., 2019). However, in vitro conditions are considerably different to those found in vivo, having, for example different nutrient availability and lack of a host immune response (Brown & Guler, 2020). Previous studies found inconsistent results in terms of whether var gene expression was impacted by culture and if it was, they disagreed on which var groups were the most affected (Zhang et al., 2011, Peters et al., 2007). Similar challenges apply to the understanding of changes in P. falciparum non-polymorphic core genes in culture, with the focus previously being on long-term laboratory adapted parasites (Claessens et al., 2017, Mackinnon et al., 2009). Consequently, direct interpretation of a short-term cultured parasite’s transcriptome remains a challenge. It is fundamental to understand var genes in context with the parasite’s core transcriptome. This could provide insights into var gene regulation and phenomena such as the proposed lower level of var gene expression in asymptomatic individuals (Almelli et al., 2014, Andrade et al., 2020).

Here we present an improved method using RNA-sequencing data for assembly, characterization, and quantification of var gene expression. This new approach overcomes previous limitations and outperforms current methods, enabling a much greater understanding of the var transcriptome. We demonstrate the power of this new approach by evaluating changes in var gene expression of paired samples from clinical isolates of P. falciparum during their early transition to in vitro culture, across several generations. We also assess changes in relation to disease phenotype and prior malaria exposure status. The use of paired samples, which are genetically identical and hence have the same var gene repertoire, allows validation of assembled transcripts and direct comparisons of expression. We complement this with a comparison of changes which occur in the non-polymorphic core transcriptome over the same transition into culture. We find a background of modest changes in var gene expression with unpredictable patterns of var gene switching, favouring an apparent convergence towards var2csa expression. More extensive changes were observed in the core transcriptome during the first cycle of culture, suggestive of a parasite stress response.
**Results**

To extend our ability to characterise var gene expression profiles and changes over time in clinical P. falciparum isolates, we set out to improve current assembly methods. Previous methods for assembling var transcripts have focussed on assembling whole transcripts (Tonkin-Hill et al., 2018, Wichers et al., 2021, Guillohon et al., 2022, Andrade et al., 2020). However, due to the diversity within PFEMP1 domains, their associations with disease severity and the fact they are not inherited together, a method that focusses on domain assembly first was also developed. In addition, a novel whole transcript approach, using a different de novo assembler, was developed and their performance compared to the method of Wichers et al. (hereafter termed “the original approach”, Figure 2). The new approaches made use of the MalariaGEN P. falciparum dataset, which led to the identification of additional multi-mapping non-core reads (a median of 3,955 reads per sample) prior to var transcript assembly (MalariaGen et al., 2021). We incorporated read error correction and improved large scaffold construction with fewer misassemblies (see Methods). We then applied this pipeline to paired ex vivo and short-term in vitro cultured parasites to enhance our understanding of the impact of short-term culturing on the var transcriptome (Figure 3). The var transcriptome was assessed at several complementary levels. First, var transcript and domain differential expression was performed. Next, changes in the dominantly expressed var gene and the homogeneity of var expression in paired samples were investigated. Finally, var group and global var gene expression changes were evaluated. This was accompanied by analysis of the core transcriptome through short-term culture.

---

**Figure 2: Novel computational pipelines for assembling var transcripts.** The original approach [green] used SoapDeNovo-Trans [k=71] to perform whole var transcript assembly. The whole transcript approach [blue] focussed on assembling whole var transcripts from the non-core reads using mraSPades [k = 71]. Contigs were then joined into longer transcripts using SSPACE. The domain approach [orange] assembled var domains first and then joined the domains into whole transcripts. Domains were assembled separately using three different de novo assemblers [SoapDeNovo2, Oases and mraSPades]. Next, a reference of assembled domains was composed and cd-hit (at sequence identity 99%) was used to remove redundant sequences. Cap3 was used to merge and extend domain assemblies. Finally, SSPACE was used to join domains together. HMM models built on the Rask et al., 2010 dataset were used to annotate the assembled transcripts. The most significant alignment was taken as the best annotation for each region of the assembled transcript (significance <= 1e-5). Identified using call-resolve-hits0. Transcripts < 500 nt were removed. A var transcript was selected if it contained at least one significantly annotated domain (in exon 1). Var transcripts that encoded only the more conserved exon 2 [ATS domain] were discarded. The three pipelines were run on the 32 malaria patient ex vivo samples from Wichers et al., 2021 (Wichers et al., 2021).
Improving var transcript assembly, annotation and quantification

A laboratory and a clinical dataset were used to assess the performance of the different var assembly pipelines (Figure 2). The laboratory dataset was a \textit{P. falciparum} 3D7 time course RNA-sequencing dataset (ENA: PRJEB31535) (Wichers \textit{et al.}, 2019). The clinical dataset contained samples from 32 adult malaria patients, hospitalised in Hamburg, Germany. 15 were malaria naïve and 17 were previously exposed to malaria. Eight of the malaria naïve patients went on to develop severe malaria and 24 had non-severe malaria (Wichers \textit{et al.}, 2021).

Our \textit{i}) new whole transcript approach, \textit{ii}) domain assembly approach, and \textit{iii}) modified version of the original approach were first applied to the \textit{P. falciparum} 3D7 time course RNA-sequencing dataset to benchmark their performance (European nucleotide archive (ENA): PRJEB31535) (Wichers \textit{et al.}, 2019) (Figure S1). The whole transcript approach performed best, achieving near perfect alignment scores for the dominantly expressed var gene. The domain and the original approach produced shorter contigs and required more contigs to assemble the var transcripts at the 8 and 16 hour post-invasion time points, when var gene expression is maximal (Figure S1c, f, g and h). However, we found high accuracies (> 0.95) across all approaches, meaning the sequences we assembled were correct (Figure S1b). The whole transcript approach also performed the best when assembling the lower expressed var genes (Figure S1e) and...
produced the fewest misassemblies compared to the original approach (14 vs 19 respectively), especially for the ring-stage samples (Table S1).

Next, the assembled transcripts produced from the original approach of Wichers et al., 2021 were compared to those produced from our new whole transcript and domain assembly approaches for ex vivo samples from German travellers. Summary statistics are shown in Table 1. The whole transcript approach produced the fewest transcripts, but of greater length than the domain approach and the original approach (Figure S2). The whole transcript approach also returned the largest N50 score, which means that it was the most contiguous assembly produced. The original approach and the new whole transcript approach gave similar results for domain expression in each sample with greater correlation in results observed between the highly expressed domains (Figure S3). Comparable results were also seen for the differentially expressed transcripts between the naive vs pre-exposed and severe vs non-severe comparisons, respectively (Figure S4). Remarkably, with the new whole transcript method, we observed a significant decrease in clearly misassembled transcripts with, for example, an N-terminal domain at an internal position (2 vs 336 misassemblies).

<table>
<thead>
<tr>
<th></th>
<th>Total no. contigs &gt;= 500nt</th>
<th>Maximum length (nt)</th>
<th>N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original approach</td>
<td>6,441</td>
<td>10,412</td>
<td>2,302</td>
</tr>
<tr>
<td>Domain approach</td>
<td>4,691</td>
<td>5,003</td>
<td>1,088</td>
</tr>
<tr>
<td>Whole transcript approach</td>
<td>3,011</td>
<td>12,586</td>
<td>5,381</td>
</tr>
</tbody>
</table>

Table 1: Statistics for the different approaches used to assemble the var transcripts. Assembly approaches were applied to malaria patient ex vivo samples (n=32) from Wichers et al., 2021 and statistics determined. Given are the total number of assembled var transcripts longer than 500 nt containing at least one significantly annotated var domain, the maximum length of the longest assembled var transcript in nucleotides and the N50 value, respectively. The N50 is defined as the sequence length of the shortest var contig, with all var contigs greater than or equal to this length together accounting for 50% of the total length of concatenated var transcript assemblies.

Overall, the new whole transcript approach performed the best on the 3D7 dataset (ENA:PRJEB31535) (Wichers et al., 2019), had the greatest N50, the longest var transcripts and produced concordant results with the original analysis on the ex vivo samples (Wichers et al., 2021). Therefore, it was selected for all subsequent analyses unless specified otherwise.

Differential expression of var transcripts from ex vivo to in vitro samples

Of the 32 clinical isolates of *P. falciparum* from the German traveller dataset, 13 underwent one replication cycle in in vitro culture, 10 of these underwent a second generation and one underwent a third generation (Table 2). Most (9/13, 69%) isolates entering culture had a single MSP1 genotype, indicative of monoclonal infections. A median of 17 million 125 bp paired-end reads, were obtained from each parasite isolate in generation 1 and generation 2, the ex vivo samples had a greater read depth than the in vitro samples (Table 2). Figure 3 shows a summary of the analysis performed.
To account for differences in parasite developmental stage within each sample, which are known to impact gene expression levels (Bozdech et al., 2003), the proportions of life cycle stages were estimated using the mixture model approach of the original analysis (Tonkin-Hill et al., 2018, Wichers et al., 2021). As a complementary approach, single cell differential composition analysis (SCDC) with the Malaria Cell Atlas as a reference was also used to determine parasite age (Dong et al., 2021, Howick et al., 2019). SCDC and the mixture model approaches produced concordant estimates that most parasites were at ring stage in all ex vivo samples and in vitro cultures (Figure S5a and 5b). Whilst there was no significant difference in ring stage proportions across the generations, we observed a slight increase in parasite age in the cultured samples. Overall, there were more rings and early trophozoites in the ex vivo samples compared to the cultured parasite samples and an increase of late trophozoite, schizont and gametocyte proportions during the culturing process (Figure S5c). The estimates produced from the mixture model approach showed high concordance with those observed by counting Giemsa stained blood smears (Figure S5d). Due to the potential confounding effect of differences in stage distribution on gene expression, we adjusted for developmental stage determined by the mixture model in all subsequent analyses.

Our new approach was applied to RNA-sequencing samples of ex vivo and short-term in vitro cultured parasites from German travellers (Wichers et al., 2021). Table S2 shows the assembled var transcripts on a per sample basis. Interestingly, we observed SSPACE did not provide improvement in terms of extending var assembled contigs in 9/37 samples. We observed a significant increase in the number of assembled var transcripts in generation 2 parasites compared to paired generation 1 parasites ($p_{adj} = 0.04$, paired Wilcoxon test). We observed no significant differences in the length of the assembled var transcripts across the generations.

Table 2: Summary of the dataset used in the analysis. RNA-sequencing was performed on 32 malaria infected German traveller samples (Wichers et al., 2021). These 32 ex vivo samples were used to compare the performance of the var assembly approaches. Parasites from 13 of these ex vivo samples underwent one cycle of in vitro replication, 10 parasite samples were also subjected to a second cycle of replication in vitro, and a single parasite isolate was also analyzed after a third cycle of replication. For the ex vivo vs short-term in vitro replication analysis only paired samples were used. The number of assembled var contigs represents results per sample using the whole transcript approach, and shows either the number of assembled var contigs significantly annotated as var gene and > = 500nt in length, or the number of assembled var transcripts identified with a length >= 1500nt and containing at least 3 significantly annotated var domains.

*PE: paired-end reads.
Three different filtering approaches were applied in comparison to maximise the likelihood that correct assemblies were taken forward for further analysis and to avoid the overinterpretation of lowly expressed partial var transcripts (Table S3). Filtering for var transcripts at least 1500 nt long and containing at least 3 significantly annotated var domains was the most restrictive, while the others required the presence of a DBLα domain within the transcript. All three filtering approaches had the same maximum length var transcript and similar N50 values. We observed a similar pattern of var transcriptional change across all three filtering approaches used (described below).

Each sample’s non-core reads, from which var assembly was performed, were mapped against a pooled reference of the assembled var transcripts from all samples, as in Wichers et al., 2021. Principal component analysis (PCA) analysis was performed on the normalised var transcript read counts (Figure 4a and b). In some patient samples, the ex vivo sample did not cluster closest to its paired generation 1 sample (patient #1, #17 and #25), suggesting there was a var transcriptional change in these samples during the first cycle of cultivation, which was not clearly related to any technical differences in culture or sequencing (explained in Methods) (Table S4). However, the rest of the samples clustered by patient identity, suggesting that transition to culture had less impact than parasite origin on their var transcriptomes. This was apparent across all three filtering approaches applied (Figure S6a and b). In line with these results, var expression homogeneity (VEH) was investigated across the samples. The calculated Bray-Curtis β diversity index showed similar clustering to that seen in the var transcriptome PCA, in which we observed general clustering by sample identity, with the exception of the same ex vivo samples previously identified (Figure 4a, Figure S7). This suggests that these patient samples undergo a greater var transcriptional change when cultured compared to the other patient samples.
Seven var transcripts were found to be significantly differentially expressed in the ex vivo vs generation 1 comparison (Figure 4c). We observed no significant differences in the expression of conserved var variants, var1-IT (p_{adj} = 0.61, paired Wilcoxon test), var1-3D7 (p_{adj} = 0.93, paired Wilcoxon test) and var2csa (p_{adj} = 0.54, paired Wilcoxon test) between paired ex vivo and generation 1 parasites. This represents a relatively small number of var transcripts differentially expressed out of the 455 assembled var transcripts.

The comparison between generation 2 and ex vivo had a smaller number of samples, however we observed a similar pattern, in terms of the differentially expressed transcripts, to that of the ex vivo vs generation 1 comparison (Figure 4d). Var2csa was significantly differentially expressed between generation 1 and generation 2 parasites (p_{adj} = 0.029, paired Wilcoxon test). No other var transcripts were significantly differentially expressed in the comparison between generation 1 vs generation 2 parasite samples (Figure 4f), suggesting the var transcriptomes of these two generations are highly similar and that any major changes in var expression occur at the start of the cultivation. The log$_2$ Fold change (logFC) observed in the ex vivo vs generation 2 comparison were greater than in the ex vivo vs generation 1 comparison, suggesting that the transcriptional changes initiated in the first cycle of cultivation, are reinforced during ongoing cultivation. However, var2csa expression appeared to decrease in some paired samples during the first cycle of cultivation (Figure 4e). Similar results were observed across all filtering approaches used (Figure S6c, d, e and f). Overall, the var transcript differential expression analysis shows relatively small changes in var expression from ex vivo to culture. These small changes might be explained by the loss of selection pressure which occurs in vivo.

Of the 7 differentially expressed transcripts in the ex vivo vs generation 1 comparison, the four downregulated var transcripts in the generation 1 samples included three encoding CD36 binding PfEMP1 proteins (CIDRα2.2-DBLα1-CIDRβ1, CIDRα2.5-DBLα1-CIDRβ1-DBLβ2, CIDRα2.10-DBLβ5-DBLβ5-DBLα1-CIDRβ1). The three transcripts upregulated in generation 1 encode the CD36-binder NTSB3-DBLα0.8-CIDRα3.4-DBLα1-CIDRβ1, the EPCR-binder DBLα2-CIDRα1.6-DBLβ12-DBLα1-CIDRβ5 and a C-terminal fragment with unknown N-terminus and associated binding phenotype (DBLγ11-DBLα1-CIDR7) (Figure 4g).

To verify that the differentially expressed transcripts were not due to differential growth of multiple parasite clones within a sample, differential expression analysis of var gene expression between ex vivo and generation 1 samples was repeated in the subset of samples with only a single MSP1 genotype (Figure S8a). There was a high correlation (Spearman’s rank correlation = 0.87) between the logFC values of these samples with those of all samples (Figure S8b). This suggests the differentially expressed var transcripts are not due to multiple Plasmodium genotypes being present or inadvertent selective culture of a single parasite clone from a polyclonal infection.

**Differential expression of var domains from ex vivo to in vitro samples**

Next, we performed differential expression analysis on the level of encoded PfEMP1 domain subtypes, as done in previous studies, which can be associated with var gene groups and receptor binding phenotypes (Tonkin-Hill et al., 2018, Wichers et al., 2021). Similar to the transcript level analysis, PCA on var domain expression (Figure 5a) showed some patients’ ex vivo samples clustering away from their respective generation 1 sample (patient #1, #2, #4, #17, #25 and #12), again indicating a greater var transcriptional change relative to the other samples during the first cycle of cultivation.
**Figure 5:** Var domain transcriptome analysis through short-term *in vitro* culture. Var transcripts for paired *ex vivo* \( n=13 \), generation 1 \( n=13 \), generation 2 \( n=10 \) and generation 3 \( n=1 \) were de novo assembled using the whole transcript approach. Var transcripts were filtered for those \( \geq 1500 \) nt in length and containing at least 3 significantly annotated var domains. Transcripts were annotated using HMM models built on the Rask et al., 2010 dataset (Rask et al., 2010). When annotating the whole transcript, the most significant alignment was taken as the best annotation for each region of the assembled transcript \( \text{e-value cut off } 1e^{-5} \). Multiple annotations were allowed on the transcript if they were not overlapping, determined using fast-resolve-hits. Var domain expression was quantified using FeatureCounts and the domain counts aggregated \( a \) PCA plot of log2 normalized read counts \( \text{adj for life cycle stage, derived from the mixture model approach} \). Points are coloured by their generation \( \text{ex vivo: purple, generation 1: red, generation 2: green and generation 3: blue} \) and labelled by their patient identity. \( b \) Volcano plot showing extent and significance of up- or down-regulation of var domain expression in *ex vivo* \( n=13 \) compared with paired generation 1 cultured parasites \( n=13 \) \( \text{red and blue, } P<0.05 \text{ after Benjamin-Hochberg adjustment for FDR; red and green, absolute log2 fold change } \log_2 \text{FC in expression } \geq 2 \). Domains with a \( \log_2 \text{FC} >2 \) represent those upregulated in generation 1 parasites. Domains with a \( \log_2 \text{FC} < -2 \) represent those downregulated in generation 1 parasites. \( c \) Volcano plot showing extent and significance of up- or down-regulation of var domain expression in *ex vivo* \( n=10 \) compared with paired generation 2 cultured parasites \( n=10 \) \( \text{green, absolute log2 fold change } \log_2 \text{FC in expression } \geq 2 \). Domains with a \( \log_2 \text{FC} >2 \) represent those upregulated in generation 2 parasites. Domains with a \( \log_2 \text{FC} < -2 \) represent those downregulated in generation 2 parasites. Differential expression analysis was performed using DESeq2 \( \text{adj for life cycle stage, derived from the mixture model approach} \).
In the generation 1 vs ex vivo comparison, a single domain was significantly differentially expressed, CIDEα2.5 associated with B-type PfEMP1 proteins and CD36-binding (Figure 5b). In the generation 2 vs ex vivo comparison, there were no domains significantly differentially expressed, however we observed large logFC values in similar domains to those changing most in the ex vivo vs generation 1 comparison (Figure 5c). No differentially expressed domains were found in the generation 1 vs generation 2 comparison.

**Cultured parasites as surrogates for assessing the in vivo var transcriptome**

Parasite isolates from malaria patients are often studied in short-term culture to investigate relationships between parasite factors such as var gene expression, cytoadhesion phenotype, and host clinical or immunological status. To understand how culture-induced changes in var gene expression may impact the interpretation of such analyses, we performed differential expression analysis comparing parasite isolates from naïve (n=6) versus previously-exposed (n=7) patients with paired ex vivo and generation 1 samples. This was performed using the ‘all sample analysis’ approach (Figure 3) with the same reference of assembled var transcripts used for the ex vivo and generation 1 analysis. There were 75 differentially expressed transcripts in the ex vivo samples, but only 18 differentially expressed transcripts in the generation 1 analysis (Figure S9a and S9b). The transcripts that were differentially expressed in the generation 1 samples were similar to those found in the ex vivo samples, but there were smaller fold change values in the generation 1 analysis. However, one transcript that was significantly upregulated in naïve patients in the ex vivo samples was found to be downregulated in the generation 1 sample analysis (encoding CD36-binding NTSB3-DBLα0.5-CIDEα2.3-DBLα1-CIDRB1). The differential selection pressure between previously exposed and malaria-naïve patients is removed during culture. Our results suggest the removal of this pressure results in the var transcriptomes of different parasites becoming more similar to each other, independent of their original in vivo transcriptional state.

**Per patient var transcript analysis**

Different parasite isolates have different var genomic repertoires, whilst differential expression analysis assumes all samples could express the same transcripts. Prior to differential expression analysis, a combined reference is created from the assemblies in all samples and read mapping is performed (Figure 3, ‘all sample analysis’). This allows reads from one sample to map to var gene assemblies from a different sample, but this mapping may not always correctly represent the true origin of each read. We sought to address these limitations by analysing var gene expression at an individual subject level, where we confirmed the MSP1 genotypes and alleles were still present after short-term in vitro cultivation. There are about 60 var genes in each *P. falciparum* genome, however, each individual parasite expresses only a single var gene at any given time. To investigate whether dominant var gene expression changes through in vitro culture, rank analysis of var transcript expression was performed (Figure S10). In most cases the dominant var transcripts were concordant between ex vivo and in vitro cultured samples from the same patient, with a single dominant var transcript. This suggests the parasites were behaving clonally, a phenomenon described previously (Bachmann *et al.*, 2011). However, we observed a change in the dominant var gene being expressed through culture in isolates from patients #6, #17 and #26. Except #17, these were not the samples identified in the transcript and domain analysis when all samples’ var gene assemblies were combined and expression quantified. Changes in the dominant var gene expression were also observed in the DBLα-tag data for these patients (described below). In patients #7, #1 and #14, the top expressed var gene remained the same, however we observed a change in the ranking of other highly expressed var genes in the cultured samples compared to the ex vivo. This suggests some parasites in these samples underwent var gene switching, and in the samples where the dominant var gene changed, a fast-switching event might have occurred (Bachmann *et al.*, 2011). Interestingly, in patient #26 we observed a switch from a dominant group A var gene to a group B and C var gene. This finding was also observed in the DBLα-tag analysis (results below). A similar finding was seen in patient #7. In the ex vivo sample, the second most expressed var transcript was a group A transcript. However, in the cultured
samples expression of this transcript was reduced and we observed an increase in the expression of group B and C var transcripts. A similar pattern was observed in the DBLα-tag analysis for patient #7, whereby the expression of a group A transcript was reduced during the first cycle of cultivation. The dominant var gene did not change in the other patient samples and the ranking of var gene expression remained similar. This again suggests that some patient samples underwent a larger var transcriptional change when cultured compared to the other patient samples, and that culturing parasites can lead to an unpredictable var transcriptional change.

In line with these results, VEH on a per patient basis showed in some patients a clear change, with the ex vivo sample diversity curve distinct from those of in vitro generation 1 and generation 2 samples (patient #4, #1 and #2) (Figure S11). Similarly, in other patient samples, we observed a clear difference in the curves of ex vivo and generation 1 samples (patient #25 and #26, both from malaria-naïve severe malaria patients). Some of these samples (#1 and #26, both from malaria-naïve severe malaria patients) also showed changes in their dominant var gene expression during culture, taken together indicating much greater var transcriptional changes in vitro compared to the other samples.

**Var group expression analysis**

A previous study found group A var genes to have a rapid transcriptional decline in culture compared to group B var genes, however another study found a decrease in both group A and group B var genes in culture (Zhang et al., 2011, Peters et al., 2007). These studies were limited as the var type was determined by analysing the sequence diversity of DBLα domains, and by quantitative PCR (qPCR) methodology which restricts analysis to quantification of known/conserved sequences. Due to these results, the expression of group A var genes vs. group B and C var genes was investigated using a paired analysis on all the DBLα (DBLα1 vs DBLα0 and DBLα2) and NTS (NTSA vs NTSB) sequences assembled from ex vivo samples and across multiple generations in culture. A linear model was created with group A expression as the explanatory variable, the generation and life cycle stage as independent variables and the patient information included as a random effect. The same was performed using group B and C expression levels.

In both approaches (DBLα and NTS) we found no significant changes in total group A or group B and C var gene expression levels (Figure 6). We observed high levels of group B and C var gene expression compared to group A in all patients, both in the ex vivo samples and the in vitro samples. In some patients we observed a decrease in group A var genes from ex vivo to generation 1 (patients #5, #2, #26, #12, #6, #9, #1 and #17)(Figure 6a), however in all but four patients (patient #1, #5, #6 and #2) the levels of group B and C var genes remained consistently high from ex vivo to generation 1 (Figure 6b). Interestingly, patients #17 and #6 also had a change in the dominant var gene expression through culture. Taken together with the preceding results, it appears that observed differences in var transcript expression occurring with transition to short-term culture are not due to modulation of recognised var classes, but due to differences in expression of particular var transcripts.
Quantification of total var gene expression

We observed a trend of decreasing total var gene expression between generations irrespective of the assembler used in the analysis (Figure S12). A similar trend is seen with the LARSFADIG count, which is commonly used as a proxy for the number of different var genes expressed (Otto et al., 2019). A linear model was created (using only paired samples from ex vivo and generation 1) [Supplementary file 1] with proportion of total gene expression dedicated to var gene expression as the explanatory variable, the generation and life cycle stage as independent variables and the patient information included as a random effect. This model showed no significant differences between generations, suggesting that differences observed in the raw data may be a consequence of small changes in developmental stage distribution in culture.
**Var expression profiling by DBLα-tag sequencing**

Deep sequencing of RT-PCR-amplified DBLα expressed sequence tags (ESTs) combined with prediction of the associated transcripts and their encoded domains using the Varia tool (Mackenzie *et al.*, 2022) was performed, as in the original analysis, to supplement the RNA-sequencing analysis. The raw Varia output file is given in Supplementary file 2. Overall, we found a high agreement between the detected DBLα-tag sequences and the *de novo* assembled var transcripts. Around 94% (median) of all unique DBLα-tag sequences detected with >10 reads were found in the RNA-sequencing approach. This is an improvement on the original approach, in which a median of 82.6% was found (Wichers *et al.*, 2021). To allow for a fair comparison of the >10 reads threshold used in the DBLα-tag approach, the upper 75th percentile of the RNA-sequencing-assembled DBLα domains were analysed. A median of 73.4% of the upper 75th percentile of the assembled DBLα domains were found in the DBLα-tag approach. This is a lower median percentage than that found in the original analysis (81.8%) and suggests the new approach might be better at capturing all expressed DBLα domains.

The new whole transcript assembly approach also had high consistency with the domain annotations predicted from Varia. Varia predicts var sequences and domain annotations based on short sequence tags, using a database of previously defined var sequences and annotations (Mackenzie *et al.*, 2022). A median of 85% was observed for the DBLα annotations identified from the DBLα-tag approach that were found in the RNA-sequencing approach. A median of 73% was observed for the DBLα-CIDR domain annotations identified from the DBLα-tag approach that were found in the RNA-sequencing approach. We also observed consistent results with the per patient analysis, in terms of changes in the dominant var gene expression (described above) (Supplementary file 2). In line with the RNA-sequencing data, the DBLα-tag approach revealed no significant differences in Group A and Group B and C groups during short-term culture, further highlighting the agreement of both methods (Figure S13).

**Core gene differential expression**

Given the modest changes in var gene expression repertoire upon culture we wanted to investigate the extent of any accompanying changes in the core parasite transcriptome. PCA was performed on core gene (*var, rif, stvor, surf* and rRNA genes removed) expression, adjusted for life cycle stage. We observed distinct clustering of *ex vivo*, generation 1, and generation 2 samples, with patient identity having much less influence (Figure 7a). There was also a change from the heterogeneity between the *ex vivo* samples to more uniform clustering of the generation 1 samples (Figure 7b), suggesting that during the first cycle of cultivation the core transcriptomes of different parasite isolates become more alike. The effect of sample type on the core gene PCA was much clearer than on the var transcript PCA (Figure 4a), suggesting the core gene transcriptome underwent a greater change relative to the var transcriptome upon transition to culture.
Figure 7: Core gene transcriptome analysis of ex vivo and short-term in vitro cultured samples. Core gene expression was assessed for paired ex vivo (n=13), generation 1 (n=13), generation 2 (n=10) and generation 3 (n=1) parasite samples. Subread align was used, as in the original analysis, to align the reads to the human genome and P. falciparum 3D7 genome, with var, erf, stwec, surf and RNA genes removed. HTSeq count was used to quantify gene counts. 

a) PCA plot of log$_2$ normalized read counts. Points are coloured by their generation (ex vivo: purple; generation 1: red; generation 2: green; and generation 3: blue) and labelled by their patient identity.

b) PCA plot of log$_2$ normalized read counts for ex vivo (purple) and paired generation 1 (red) samples. Points are labelled by their patient identity.

c) Volcano plot showing extent and significance of up- or down-regulation of core gene expression in ex vivo (n=13) compared with paired generation 1 cultured parasites (n=13) (red and blue: \( P < 0.05 \) after Benjamini-Hochberg adjustment for FDR; red and green: absolute log$_2$ fold change \( \log_2 FC \) in expression \( \geq 2 \)). Genes with a \( \log_2 FC > 2 \) represent those upregulated in generation 1 parasites. Genes with a \( \log_2 FC < -2 \) represent those downregulated in generation 1 parasites.

d) Volcano plot showing extent and significance of up- or down-regulation of core gene expression in ex vivo (n=10) compared with paired generation 2 cultured parasites (n=10) (red: \( P < 0.05 \) after Benjamini-Hochberg adjustment for FDR; red and green: absolute log$_2$ fold change \( \log_2 FC \) in expression \( \geq 2 \)). Genes with a \( \log_2 FC > 2 \) represent those upregulated in generation 2 parasites. Genes with a \( \log_2 FC < -2 \) represent those downregulated in generation 2 parasites. Normalized read counts of the core gene analysis were adjusted for life cycle stage, derived from the mixture model approach.
In total, 920 core genes (19% of the core transcriptome) were found to be differentially expressed after adjusting for life cycle stages using the mixture model approach between ex vivo and generation 1 samples (Supplementary file 3). The majority were upregulated, indicating a substantial transcriptional change during the first cycle of in vitro cultivation (Figure 7c). 74 genes were found to be upregulated in generation 2 when compared to the ex vivo samples, many with log₂FC greater than those in the ex vivo vs generation 1 comparison (Figure 7d). No genes were found to be significantly differentially expressed between generation 1 and generation 2. However, five genes had a log₂FC >= 2 and were all upregulated in generation 2 compared to generation 1. Interestingly, the gene with the greatest fold change, encoding ROM3 (PF3D7_0828000), was also found to be significantly downregulated in generation 1 parasites in the ex vivo vs generation 1 analysis. The other four genes were also found to be non-significantly downregulated in generation 1 parasites in the ex vivo vs generation 1 analysis. Similar to the var transcriptome analysis, this suggests changes in gene expression during the first cycle of cultivation are the greatest compared to the other cycles.

The most significantly upregulated genes (in terms of fold change) in generation 1 contained several small nuclear RNAs, spliceosomal RNAs and non-coding RNAs (ncRNAs). 16 ncRNAs were found upregulated in generation 1, with several RNA-associated proteins having large fold changes (log₂FC > 7) Significant gene ontology (GO) terms and Kyoto encyclopedia of genes and genomes (KEGG) pathways for the core genes upregulated in generation 1 included ‘entry into host’, ‘movement into host’ and ‘cytoskeletal organisation’ suggesting the parasites undergo a change in invasion efficiency, which is connected to the cytoskeleton, during their first cycle of in vitro cultivation (Figure S14). We observed eight AP2 transcription factors upregulated in generation 1 (PF3D7_0404100/AP2-SP2, PF3D7_0604100/SIP2, PF3D7_0611200/AP2-EXP2, PF3D7_0613800, PF3D7_0802100/AP2-LT, PF3D7_1143100/AP2-O, PF3D7_1239200, PF3D7_1456000/AP2-HC) with no AP2 transcription factors found to be downregulated in generation 1. To confirm the core gene expression changes identified were not due to the increase in parasite age during culture, as indicated by upregulation of many schizont-related genes, core gene differential expression analysis was performed on paired ex vivo and generation 1 samples that contained no schizonts or gametocytes in generation 1. The same genes were identified as significantly differentially expressed with a Spearman’s rank correlation of 0.99 for the log₂FC correlation between this restricted sample approach and those produced using all samples (Figure S15).

**Cultured parasites as surrogates for assessing the in vivo core gene transcriptome**

Due to the high number of core genes found to be significantly differentially expressed in the ex vivo vs generation 1 comparison, we further investigated the suitability of cultured parasites being used as surrogates for assessing the in vivo core gene transcriptome. We performed differential expression analysis comparing parasite isolates from naïve (n=6) vs pre-exposed (n=7) patients, between their ex vivo samples and then separately between their paired generation 1 samples. Interestingly, when using the ex vivo samples, we observed 206 core genes significantly upregulated in naïve patients compared to pre-exposed patients (Figure S16a). Conversely, we observed no differentially expressed genes in the naïve vs pre-exposed analysis of the paired generation 1 samples (Figure S16b). Taken together, this suggests one cycle of cultivation shifts the core transcriptomes of parasites to be more alike each other, diminishing inferences about parasite biology in vivo.

**Discussion**

Multiple lines of evidence point to PfEMP1 as a major determinant of malaria pathogenesis, but previous approaches for characterising var expression profiles in field samples have limited in vivo studies of PfEMP1 function, regulation, and association with clinical symptoms (Tarr et al., 2018, Lee et al., 2018, Warimwe et al., 2013, Rorick et al., 2013, Zhang et al., 2011, Taylor et al., 2002). A more recent approach, based on RNA-sequencing, overcame many of the limitations imposed by the previous primer-based
methods (Tonkin-Hill et al., 2018, Wichers et al., 2021). However, depending on the expression level and sequencing depth, var transcripts were found to be fragmented and only a partial reconstruction of the var transcriptome was achieved (Tonkin-Hill et al., 2018, Wichers et al., 2021, Andrade et al., 2020, Guillochon et al., 2022, Yamagishi et al., 2014). The present study developed a novel approach for var gene assembly and quantification that overcomes many of these limitations.

Our new approach used the most geographically diverse reference of var gene sequences to date, which improved the identification of reads derived from var transcripts. This is crucial when analysing patient samples with low parasitaemia where var transcripts are hard to assemble due to their low abundance (Guillochon et al., 2022). Our approach has wide utility due to stable performance on both laboratory-adapted and clinical samples. Concordance in the different var expression profiling approaches (RNA-sequencing and DBLα-tag) increased using the new approach by 11.4%, when compared to the original approach (94% in the whole transcript approach compared to 82.6% in Wichers et al., 2021. This suggests the new approach provides a more reliable and robust method for characterising var genes.

Having a low number of long contigs is desirable in any de novo assembly. This reflects a continuous assembly, as opposed to a highly fragmented one where polymorphic and repeat regions could not be resolved (Lischer & Shimizu, 2017). An excessive number of contigs cannot be reasonably handled computationally and results from a high level of ambiguity in the assembly (Yang et al., 2012). We observed a greater than 50% reduction in the number of contigs produced in our new approach, which also had a 21% increase in the maximum length of the assembled var transcripts, when compared to the original approach. It doubled the assembly continuity and assembled an average 13% more of the var transcripts. This was particularly apparent in the N-terminal region, which has often been poorly characterised by existing approaches. The original approach failed to assemble the N-terminal region in 58% of the samples, compared to just 4% in the new approach with assembly consistently achieved with an accuracy > 90%. This is important because the N-terminal region is known to contribute to the adhesion phenotypes of most PfEMP1s.

The new approach allows for var transcript reconstruction across a range of expression levels, which is required when characterising var transcripts from multi-clonal infections. Assembly completeness of the lowly expressed var genes increased five-fold using the new approach. Biases towards certain parasite stages have been observed in non-severe and severe malaria cases, so it is valuable to assemble the var transcripts from different life cycle stages (Tonkin-Hill et al., 2018). Our new approach is not limited by parasite stage. It was able to assemble the whole var transcript, in a single contig, at later stages in the P. falciparum 3D7 intra-erythrocytic cycle, something previously unachievable. The new approach allows for a more accurate and complete picture of the var transcriptome. It provides new perspectives for relating var expression to regulation, co-expression, epigenetics and malaria pathogenesis. It has huge applicability, for example analysis of patient samples with different clinical outcomes and longitudinal tracking of infections in vivo. It represents a crucial improvement for quantifying the var transcriptome. In this work, the improved approach for var gene assembly and quantification was used to characterise var gene expression during transition from in vivo to short-term culture.

Studies have been performed investigating differences between long-term laboratory-adapted clones and clinical isolates, with hundreds of genes found to be differentially expressed (Hoo et al., 2019, Tarr et al., 2018, Mackinnon et al., 2009). Surprisingly, studies investigating the impact of short-term culture on parasites are extremely limited, despite it being commonly undertaken for making inferences about the in vivo transcriptome (Vignali et al., 2011). Using the new var assembly approach, we found that var gene expression remains relatively stable during transition to culture, with only a small number of differentially expressed transcripts across the whole dataset. These included var2csa, which had increased expression from generation 1 to generation 2. It has previously been suggested that long-term cultured parasites
converge to expressing var2csa, but our findings suggest this begins within two cycles of cultivation (Zhang et al., 2022, Mok et al., 2008). Switching to var2csa has been shown to be favourable and suggests the default var gene of choice upon perturbation to var specific heterochromatin (Ukaegbu et al., 2015). These studies also suggested var2csa has a unique role in var gene switching and our results are consistent with the role of var2csa as the dominant ‘sink node’ (Zhang et al., 2022, Ukaegbu et al., 2015, Ukaegbu et al., 2014, Mok et al., 2008). A previous study suggested in vitro cultivation of controlled human malaria infection samples resulted in dramatic changes in var gene expression (Lavstsen et al., 2005, Peters et al., 2007). Some individuals in our analysis showed more pronounced and unpredictable changes. In these individuals, the dominant var gene being expressed changed within one cycle of cultivation. This implies short-term culture can result in unpredictable var gene expression as observed previously using a semi-quantitative RT-PCR approach (Bachmann et al., 2011) and that one would need to confirm in vivo expression matches in vitro expression. This can be achieved using the assembly approach described here.

We observed no generalised pattern of up- or downregulation of specific var groups following transition to culture. This implies there is probably not a selection event occurring during culture but may represent a loss of selection that is present in vivo. A global down regulation of certain var groups might only occur as a selective process over many cycles in extended culture. Determining changes in var group expression levels are difficult using degenerate qPCR primers bias and previous studies have found conflicting results in terms of changes of expression of var groups through cultivation. Zhang et al., 2011 found a rapid transcriptional decline of group A and group B var genes, however Peters et al., 2007 found group A var genes to have a high rate of downregulation, when compared to group B var genes. These studies differed in the stage distribution of the parasites and were limited in measuring enough variants through their use of primers. Our new approach allowed for the identification of more sequences, with 26.6% of assembled DBLα domains not found via the DBLα-tag approach. This better coverage of the expressed var diversity was not possible in these previous studies and may explain discrepancies observed.

Generally, there was a high consensus between all levels of var gene analysis and changes observed during short-term in vitro cultivation. However, the impact of short-term culture was the most apparent at the var transcript level and became less clear at higher levels. This highlights the need for accurate characterisation of full length var transcripts and analysis of the var transcriptome at different levels, both of which can be achieved with the new approach developed here.

We saw striking changes in the core gene transcriptomes between ex vivo and generation 1 parasites with 19% of the core genome being differentially expressed. A previous study showed that expression of 18% of core genes were significantly altered after ~50 cycles through culture (Mackinnon et al., 2009), but our data suggest that much of this change occurs early in the transition to culture. We observed genes with functions unrelated to ring-stage parasites were among those most significantly expressed in the generation 1 vs ex vivo analysis, suggesting the culture conditions may temporarily dysregulate stage-specific expression patterns or result in the parasites undergoing a rapid adaptation response (Andreadaki et al., 2020, Beeson et al., 2016). Several AP2 transcription factors (AP2-SP2, AP2-EXP2, AP2-LT, AP2-O and AP2-HC) were upregulated in generation 1. AP2-HC has been shown to be expressed in asexual parasites (Carrington et al., 2021). AP2-O is thought to be specific for the ookinete stage and AP2-SP2 plays a key role in sporozoite stage specific gene expression (Kaneko et al., 2015, Yuda et al., 2010). Our findings are consistent with another study investigating the impact of long-term culture (Mackinnon et al., 2009) which also found genes like merozoite surface proteins differentially expressed, however they were downregulated in long-term cultured parasites, whereas we found them upregulated in generation 1. This suggests short-term cultured parasites might be transcriptionally different from long-term cultured parasites, especially in their invasion capabilities, something previously unobserved. Several genes involved in the stress response of parasites were upregulated in generation 1, for example Dnaj proteins, serine
proteases and ATP dependent CLP proteases (Oakley et al., 2007). The similarity of the core transcriptomes of the *in vitro* samples compared to the heterogeneity seen in the *ex vivo* samples could be explained by a stress response upon entry to culture. Studies investigating whether the dysregulation of stage specific expression and the expression of stress associated genes persist in long-term culture are required to understand whether they are important for growth in culture. Critically, the marked differences presented here suggest the impact of short-term culture can override differences observed in both the *in vivo* core and *var* transcriptomes of different disease manifestations.

In summary, we present an enhanced approach for *var* transcript assembly which allows for *var* gene expression to be studied in connection to *P. falciparum*’s core transcriptome through RNA-sequencing. This will be useful for expanding our understanding of *var* gene regulation and function in *in vivo* samples. As an example of the capabilities of the new approach, the method was used to quantify differences in gene expression upon short-term culture adaptation. This revealed that inferences from clinical isolates of *P. falciparum* put into short-term culture must be made with a degree of caution. Whilst *var* gene expression is often maintained, unpredictable switching does occur, necessitating that the similarity of *in vivo* and *in vitro* expression should be confirmed. The more extreme changes in the core transcriptome could have much bigger implications for understanding other aspects of parasite biology such as growth rates and drug susceptibility and raise a need for additional caution. Further work is needed to examine *var* and core transcriptome changes during longer term culture on a larger sample size. Understanding the ground truth of the *var* expression repertoire of *Plasmodium* field isolates still presents a unique challenge and this work expands the database of *var* sequences globally. The increase in long-read sequencing and the growing size of *var* gene databases containing isolates from across the globe will help overcome this issue in future studies.

**Materials and Methods**

**Ethics statement**

The study was conducted according to the principles of the Declaration of Helsinki, 6th edition, and the International Conference on Harmonization-Good Clinical Practice (ICH-GCP) guidelines. All 32 patients were treated as inpatients or outpatients in Hamburg, Germany (outpatient clinic of the University Medical Center Hamburg-Eppendorf (UKE) at the Bernhard Nocht Institute for Tropical Medicine, UKE, Bundeswehrkrankenhaus) (Wichers et al., 2021). Blood samples for this analysis were collected after patients had been informed about the aims and risks of the study and had signed an informed consent form for voluntary blood collection (n=21). In the remaining cases, no intended blood samples were collected but residuals from diagnostic blood samples were used (n=11). The study was approved by the responsible ethics committee (Ethics Committee of the Hamburg Medical Association, reference numbers PV3828 and PV4539).

**Blood sampling, processing and *in vitro* cultivation of *P. falciparum***

EDTA blood samples (1–30 mL) were collected from 32 adult falciparum malaria patients for *ex vivo* transcriptome profiling as reported by Wichers et al., 2021 (Wichers et al., 2021), hereafter termed ‘the original analysis’. Erythrocytes were isolated by Ficoll gradient centrifugation, followed by filtration through Plasmodipur filters (EuroProxima) to remove residual granulocytes. At least 400 µl of the purified erythrocytes were quickly lysed in 5 volumes of pre-warmed TRIzol (ThermoFisher Scientific) and stored at -80°C until further processing (“*ex vivo* samples”). When available, the remainder was then transferred to *in vitro* culture either without the addition of allogeneic red cells or with the addition of O+ human red cells (blood bank, UKE) for dilution according to a protocol adopted from Trager and Jensen (Table S4).

Cultures were maintained at 37°C in an atmosphere of 1% O2, 5% CO2, and 94% N2 using RPMI complete medium containing 10% heat-inactivated human serum (A+, Interstate Blood Bank, Inc., Memphis, USA).
Cultures were sampled for RNA purification at the ring stage by microscopic observation of the individual growth of parasite isolates, and harvesting was performed at the appropriate time without prior synchronization treatment (“in vitro samples”). 13 of these ex vivo samples underwent one cycle of in vitro cultivation, ten of these generation 1 samples underwent a second cycle of in vitro cultivation. One of these generation 2 samples underwent a third cycle of in vitro cultivation (Table 1). In addition, an aliquot of ex vivo erythrocytes (approximately 50–100 μl) and aliquots of in vitro cell cultures collected as indicated in Supplementary file 4 were processed for gDNA purification and MSP1 genotyping as described elsewhere (Wichers et al., 2021, Robert et al., 1996).

RNA purification, RNA-sequencing library preparation, and sequencing
RNA purification was performed as described in Wichers et al., 2021, using TRizol in combination with the RNeasy MinElute Kit (Qiagen) and DNase digestion (DNase I, Qiagen). Human globin mRNA was depleted from all samples except from samples #1 and #2 using the GLOBINclear kit (ThermoFisher Scientific). The median RIN value over all ex vivo samples was 6.75 (IQR: 5.93–7.40), although this measurement has only limited significance for samples containing RNA of two species. Accordingly, the RIN value increased upon cultivation for all in vitro samples (Supplementary file 5). Customized library construction in accordance to Tonkin-Hill et al., 2018, including amplification with KAPA polymerase and HiSeq 2500 125 bp paired-end sequencing was performed by BGI Genomics Co. (Hong Kong).

Methods for assembling var genes
Previously Oases, Velvet, SoapDeNovo-Trans or MaSuRCA have been used for var transcript assembly (Wichers et al., 2021, Andrade et al., 2020, Otto et al., 2019, Tonkin-Hill et al., 2018). Previous methods either did not incorporate read error correction or focussed on gene assembly, as opposed to transcript assembly (Schulz et al., 2012, Zerbino & Birney, 2008, Xie et al., 2014, Zimin et al., 2013). Read error correction is important for var transcript assembly due to the highly repetitive nature of the P. falciparum genome. Recent methods have also focused on whole transcript assembly, as opposed to initial separate domain assembly followed by transcript assembly (Wichers et al., 2021, Andrade et al., 2020, Otto et al., 2019, Tonkin-Hill et al., 2018). The original analysis used SoapDeNovo-Trans to assemble the var transcripts, however it is currently not possible to run all steps in the original approach, due to certain tools being improved and updated. Therefore, SoapDeNovo-Trans (k=71) was used and termed the original approach.

Here, two novel methods for whole var transcript and var domain assembly were developed and their performance was evaluated in comparison to the original approach (Figure 2b). In both methods the reads were first mapped to the human g38 genome and any mapped reads were removed. Next, the unmapped reads were mapped to a modified P. falciparum 3D7 genome with var genes removed, to identify multi-mapping reads commonly present in Plasmodium RNA-sequencing datasets. Any mapped reads were removed. In parallel, the unmapped RNA reads from the human mapping stage were mapped against a reference of field isolate var exon 1 sequences and the mapped reads identified (Otto et al., 2019). These reads were combined with the unmapped reads from the 3D7 genome mapping stage and duplicate reads removed. All mapping was performed using sub-read align as in the original analysis (Wichers et al., 2021). The reads identified at the end of this process are referred to as “non-core reads”.

Whole var transcript and var domain assembly methods
For whole var transcript assembly the non-core reads, for each sample separately, were assembled using rnaSPAdes (k-mer =71, read_error_correction on) (Bushmanova et al., 2019). Contigs were joined into larger scaffolds using SSPACE (parameters -n 31 -x 0 -k 10) (Boetzer et al., 2011). Transcripts < 500nt were excluded, as in the original approach. The included transcripts were annotated using hidden Markov models (HMM) (Finn et al., 2011) built on the Rask et al., 2010 dataset and used in Tonkin-Hill et al., 2018.
When annotating the whole transcript, the most significant annotation was taken as the best annotation for each region of the assembled transcript (e-value cut off 1e-5). Multiple annotations were allowed on the transcript if they were not overlapping, determined using cath-resolve-hits (Lewis et al., 2019). Scripts are available in the GitHub repository (https://github.com/ClareAndradiBrown/varAssembly).

In the var domain assembly approach, separate domains were assembled first and then joined up to form transcripts. First, the non-core reads were mapped (nucleotide basic local alignment tool (blastn) short read option) to the domain sequences as defined in Rask et al., 2010. This was found to produce similar results when compared to using tblastx. An e-value threshold of 1e-30 was used for the more conserved DBLa domains and an e-value of 1e-10 for the other domains. Next, the reads mapping to the different domains were assembled separately. rnaSPAdes (read_error_correction on, k-mer = 15), Oases (kmer = 15) and SoapDeNovo2 (kmer = 15) were all used to assemble the reads separately (Bushmanova et al., 2019, Xie et al., 2014, Schulz et al., 2012). The output of the different assemblers was combined into a per sample reference of domain sequences. Redundancy was removed in the reference using cd-hit (-n 8-c 0.99) (at sequence identity = 99%) (Fu et al., 2012). Cap3 was used to merge and extend the domain assemblies (Huang & Madan, 1999). SSPACE was used to join the domains together (parameters -n 31 -x 0 -k 10) (Boetzer et al., 2011). Transcript annotation was performed as in the whole transcript approach, with transcripts < 500 nt removed. Significantly annotated (1e-5) transcripts were identified and selected. The most significant annotation was selected as the best annotation for each region, with multiple annotations allowed on a single transcript if the regions were not overlapping. For both methods, a var transcript was selected if it contained at least one significantly annotated domain (in exon 1). Var transcripts that encoded only the more conserved exon 2 (acidic terminal segment (ATS) domain) were discarded.

Validation on RNA-sequencing dataset from P. falciparum reference strain 3D7

Both new approaches and the original approach (SoapDeNovo-Trans, k = 71) (Wichers et al., 2021, Tonkin-Hill et al., 2018) were run on a public RNA-sequencing dataset of the intra-erythrocytic life cycle stages of cultured P. falciparum 3D7 strain, sampled at 8-hr intervals up until 40 hrs post infection and then at 4 hr intervals up until 48 hrs post infection (ENA: PRJEB31535) (Wichers et al., 2019). This provided a validation of all three approaches due to the true sequence of the var genes being known in P. falciparum 3D7 strain. Therefore, we compared the assembled sequences from all three approaches to the true sequence. The first best hit (significance threshold = 1e-10) was chosen for each contig. The alignment score was used to evaluate the performance of each method. The alignment score represents $\sqrt{\text{accuracy} \times \text{recovery}}$. The accuracy is the proportion of bases that are correct in the assembled transcript and the recovery reflects what proportion of the true transcript was assembled. Misassemblies were counted as transcripts that had a percentage identity < 99% to their best hit (i.e. the var transcript is not 100% contained against the reference).

Comparison of approaches for var assembly on ex vivo samples

The var transcripts assembled from the 32 ex vivo samples using the original approach were compared to those produced from the whole transcript and domain assembly approaches. The whole transcript approach was chosen for subsequent analysis and all assembled var transcripts from this approach were combined into a reference, as in the original method (Wichers et al., 2021).

Removal of var transcripts with sequence id >= 99% prior to mapping was not performed in the original analysis. To overcome this, var transcripts were removed if they had a sequence id >= 99% against the full complement in the whole transcript approach, using cd-hit-est (Fu et al., 2012). Removing redundancy in the reference of assembled var transcripts across all samples led to the removal of 1,316 assembled contigs generated from the whole transcript approach.

This reference then represented all assembled var transcripts across all samples in the given analysis. The non-core reads were mapped against this reference and quantification was performed using Salmon (Patro...
et al., 2017). DESeq2 was used to perform differential expression analysis between severe versus non-severe groups and naïve versus pre-exposed groups (Love et al., 2014).

The assembled var transcripts produced by the whole transcript assembly approach had their expression quantified at the transcript and domain level, as in the original method, and the results were compared to those obtained by the original method. To quantify domain expression, featureCounts was used, as in the original method with the counts for each domain aggregated (Liao et al., 2014). Correlation analysis between the domain’s counts from the whole transcript approach and the original method was performed for each ex vivo sample. Differential expression analysis was also performed using DESeq2, as in the original analysis and the results compared (Love et al., 2014, Wichers et al., 2021).

Estimation of parasite lifecycle stage distribution in ex vivo and short-term in vitro samples

To determine the parasite life cycle stage proportions for each sample the mixture model approach of the original analysis (Tonkin-Hill et al., 2018, Wichers et al., 2021) and the SCDC approach were used (Dong et al., 2021, Howick et al., 2019). Recently, it has been determined that species-agnostic reference datasets can be used for efficient and accurate gene expression deconvolution of bulk RNA-sequencing data from any Plasmodium species and for correct gene expression analyses for biases caused by differences in stage composition among samples (Tebben et al., 2022). Therefore, the Plasmodium berghei single cell atlas was used as reference with restriction to 1:1 orthologs between P. berghei and P. falciparum. This reference was chosen as it contained reference transcriptomes for the gametocyte stage. To ensure consistency with the original analysis, proportions from the mixture model approach were used for all subsequent analyses (Wichers et al., 2021). For comparison, the proportion of different stages of the parasite life cycle in the ex vivo and in vitro samples was determined by two independent readers in Giemsa-stained thin blood smears. The same classification as the mixture model approach was used (8, 19, 30, and 42 hours post infection corresponding to ring, early trophozoite, late trophozoite and schizont stages respectively).

Significant differences in ring stage proportions were tested using pairwise Wilcoxon tests. For the other stages, a modified Wilcoxon rank test for zero-inflated data was used (Wang et al., 2021).

Differential expression of var transcripts and domains from ex vivo to in vitro samples

The whole transcript approach was applied to the paired ex vivo and in vitro samples. Significant differences in the number of assembled var transcripts and the length of the transcripts across the generations was tested using the paired Wilcoxon test.

Redundancy was removed from the assembled var transcripts and transcripts and domains were quantified using the approach described above. Three additional filtering steps were applied separately to this reference of assembled var transcripts to ensure the var transcripts that went on to have their expression quantified represented true var transcripts. The first method restricted var transcripts to those greater than 1500nt containing at least 3 significantly annotated var domains, one of which had to be a DBLα domain. The second restricted var transcripts to those greater than 1500nt and containing a DBLα domain. The third approach restricted var transcripts to those greater than 1500nt with at least 3 significant var domain annotations. Domain expression was quantified using featureCounts, as described above (Liao et al., 2014).

DESeq2 was used to test for differential transcript and domain expression, with five expected read counts in at least three patient isolates required, with life cycle stage and patient identity used as covariates. For the ex vivo versus in vitro comparisons, only ex vivo samples that had paired samples in generation 1 underwent differential expression analysis, given the extreme nature of the polymorphism seen in the var genes. Differential expression analysis was performed for all three filtering methods described above. Differential expression analysis was also performed using samples that had a single MSP1 genotype, to ensure differences observed were not a consequence of having multiple Plasmodium genotypes present (Love et al., 2014).
To check for the differential expression of conserved var gene variants var1-3D7, var1-IT and var2csa, all assembled transcripts significantly annotated as such were identified. For each conserved gene, Salmon normalised read counts (adjusted for life cycle stage) were summed and expression compared across the generations using a pairwise Wilcoxon rank test.

**Per patient var transcript expression**

A limitation of var transcript differential expression analysis is that it assumes all var sequences have the possibility of being expressed in all samples. However, since each parasite isolate has a different set of var gene sequences, this assumption is not completely valid. To account for this, var transcript expression analysis was performed on a per patient basis. For each patient, the paired ex vivo and in vitro samples were analysed. The assembled var transcripts (at least 1500nt and containing 3 significantly annotated var domains) across all the generations for a patient were combined into a reference, redundancy was removed as described above, and expression was quantified using Salmon (Patro et al., 2017). Var transcript expression was ranked, and the rankings compared across the generations.

**Var expression homogeneity (VEH)**

VEH is defined as the extent to which a small number of var gene sequences dominate an isolate’s expression profile (Warimwe et al., 2013). Previously, this has been evaluated by calculating a commonly used α diversity index, the Simpson’s index of diversity. Different α diversity indexes put different weights on evenness and richness. To overcome the issue of choosing one metric, α diversity curves were calculated (Wagner et al., 2018). Equation 1 is the computational formula for diversity curves. $D$ is calculated for $q$ in the range 0 to 3 with a step increase of 0.1 and $p$ in this analysis represented the proportion of var gene expression dedicated to var transcript $k$. $q$ determined how much weight is given to rare vs abundant var transcripts. The smaller the $q$ value, the less weight was given to the more abundant var transcript. VEH was investigated on a per patient basis. The Bray-Curtis β diversity index was also calculated across all samples. Bray-Curtis measures the dissimilarity between samples and was chosen for its ability to handle absences of certain transcripts (Jari Oksanen, 2018).

\[ D(q) = \left( \sum_{k=1}^{K} p_k^q \right)^{\frac{1}{1-q}} \]

**Equation 1**

**Var group expression analysis**

The type of the var gene is determined by multiple parameters: upstream sequence (ups), chromosomal location, direction of transcription and domain composition. All regular var genes encode a DBLα domain in the N-terminus of the PfEMP1 protein (Figure 1c). The type of this domain correlates with previously defined var gene groups, with group A encoding DBLα1, groups B and C encoding DBLα0 and group B encoding a DBLα2 (chimera between DBLα0 and DBLα1)(Figure 1c). The DBLα domain sequence for each transcript was determined and for each patient a reference of all assembled DBLα domains combined. The relevant sample’s non-core reads were mapped to this using Salmon and DBLα expression quantified (Patro et al., 2017). DESeq2 normalisation was performed, with patient identity and life cycle stage proportions included as covariates and differences in the amounts of var transcripts of group A compared with groups B and C assessed (Love et al., 2014). A similar approach was repeated for NTS domains. NTSα domains are found encoded in group A var genes and NTSB domains are found encoded in group B and C var genes (Figure 1c).

**Quantification of total var gene expression**
The RNA-sequencing reads were blastn (with the short-blastn option on and significance = 1e-10) against the LARSFADIG nucleotide sequences (142 unique LARSFADIG sequences) to identify reads containing the LARSFADIG motifs. This approach has been described previously (Andrade et al., 2020). Once the reads containing the LARSFADIG motifs had been identified, they were used to assemble the LARSFADIG motif. Trinity (Henschel, 2012) and rnaSPAdes (Bushanova et al., 2019) were used separately to assemble the LARSFADIG motif, and the results compared. The sequencing reads were mapped back against the assemblies using bwa mem (Li, 2013), parameter -k 31 -a (as in Andrade et al., 2020). Coverage over the LARSFADIG motif was assessed by determining the coverage over the middle of the motif (S) using Samtools depth (Danecek et al., 2021). These values were divided by the number of reads mapped to the var exon 1 database and the 3D7 genome (which had var genes removed) to represent the proportion of total gene expression dedicated to var gene expression (similar to an RPKM). The results of both approaches were compared. This method has been validated on 3D7, IT and HB3 Plasmodium strains. Significant differences in total var gene expression were tested by constructing a linear model with the proportion of gene expression dedicated to var gene expression as the explanatory variable, the generation and life cycle stage as an independent variables and the patient identity included as a random effect.

**Var expression profiling by DBLα-tag sequencing**

DBLα-tag sequence analysis was performed as in the original analysis (Wichers et al., 2021), with Varia used to predict domain composition (Mackenzie et al., 2022). The proportion of transcripts encoding NTSA, NTSB, DBLα1, DBLα2 and DBLα0 domains were determined for each sample. These expression levels were used as an alternative approach to see whether there were changes in the var group expression levels through culture.

The consistency of domain annotations was also investigated between the DBLα-tag approach and the assembled transcripts. This was investigated on a per patient basis, with all the predicted annotations from the DBLα-tag approach for a given patient combined. These were compared to the annotations from all assembled transcripts for a given patient. DBLα annotations and DBLα-CIDR annotations were compared.

For comparison of both approaches (DBLα-tag sequencing and our new whole transcript approach), the same analysis was performed as in the original analysis (Wichers et al., 2021). All conserved variants (var1, var2csa and var3) were removed as they were not properly amplified by the DBLα-tag approach. To identify how many assembled transcripts, specifically the DBLα region, were found in the DBLα-tag approach, we applied BLAST. As in the original analysis, a BLAST database was created from the DBLα-tag cluster results and screened for the occurrence of those assembled DBLα regions with more than 97% seq id using the ‘megablast’ option. This was restricted to the assembled DBLα regions that were expressed in the top 75th percentile to allow for a fair comparison, as only DBLα-tag clusters with more than 10 reads were considered. Similarly, to identify how many DBLα-tag sequences were found in the assembled transcripts, a BLAST database was created from the assembled transcripts and screened for the occurrence of the DBLα-tag sequences with more than 97% seq id using the ‘megablast’ option. This was performed for each sample.

**Core gene differential expression analysis**

Subread align was used, as in the original analysis, to align the reads to the human genome and *P. falciparum* 3D7 genome, with var, rif, stevor, surf and rRNA genes removed (Liao et al., 2013). HTSeq count was used to quantify gene counts (Anders et al., 2015). DESeq2 was used to test for differentially expressed genes with five read counts in at least three samples being required (Love et al., 2014). Parasite life cycle stages and patient identity were included as covariates. GO and KEGG analysis was performed
using ShinyGo and significant terms were defined by having a Bonferroni corrected p-value < 0.05 (Ge et al., 2020).

**Funding / Acknowledgements**

CAB received support from the Wellcome Trust (4-Year PhD programme, grant number 220123/Z/20/Z). Infrastructure support for this research was provided by the NIHR Imperial Biomedical Research Centre and Imperial College Research Computing Service, DOI: 10.14469/hpc/2232.

JSWM, YDH and AB were funded by the German Research Foundation (DFG) grants BA 5213/3-1 (project #323750012) and BA 5213/6-1 (project #433302244).

TO is supported by the Wellcome Trust grant 104111/Z/14/ZR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

JB acknowledges support from Wellcome (100993/Z/13/Z)

**Author contribution**

Conceptualization: CAB, TDO, AB, AJC

Methodology: CAB, MFD, TL, TDO

Software: CAB

Validation: CAB, AB

Formal analysis: CAB

Investigation: JSW, HvT, YDH, JAMS, HSH, EFH, AB

Resources: TL, AJC, AB

Data curation: CAB, AB

Writing – original draft: CAB, TDO, AJC, AB

Writing – review & editing: CAB, JSW, MFD, TL, TWG, JB, TDO, AJC, AB

Visualization: CAB

Supervision: TWG, JB, TDO, AJC, AB

Project Administration: CAB, AJC, AB

Funding acquisition: AJC, AB

All authors read and approved the manuscript.

**Competing interests**

No competing interests declared.

**References**


