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-Brown[1,2,3], Jan Stephan Wichers-Misterek[4-6], Heidrun von Thien[4-6], Yannick D. Höppner[4-6], Judith A. M. Scholz[4], Helle Hansson[7,8], Emma Filtenborg Hocke[7,8], Tim-Wolf Gilberger[4-6], Michael F. Duffy[9], Thomas Lavstsen[7,8], Jake Baum[2,10], Thomas D. Otto[11*], Aubrey J. Cunnington[1,3*¥], Anna Bachmann[4-6,12*¥]

[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]Center for Medical Parasitology, Department of Immunology and Microbiology, University of Copenhagen, 2200 Copenhagen, Denmark
[8]Department of Infectious Diseases, Copenhagen University Hospital, 2200 Copenhagen, Denmark
[9]Department of Microbiology and Immunology, University of Melbourne, Melbourne/Parkville VIC 3052, Australia
[10]School of Biomedical Sciences, Faculty of Medicine & Health, UNSW, Kensington, Sydney, 2052, Australia
[11]School of Infection & Immunity, MVLS, University of Glasgow, UK
[12]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 outperforms the earlier approach of Wichers *et al*., 2021 on both laboratory and clinical isolates across a combination of metrics. Importantly, the tool can 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 of parasite isolates 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. Modest but 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, but our efforts to combat the disease are challenged by its complex life cycle and its sophisticated immune evasion strategies. *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 polymorphic *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).

Despite their sequence polymorphism, *var* genes could be classified into four categories (A, B, C, and E) according to their chromosomal location, transcriptional direction, type of 5'-upstream sequence (UPSA–E), and encoded protein domains with associated binding phenotype (Figure 1) (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 proteins exhibit a varying but semi-ordered composition of additional DBL and CIDR domains of different subtypes (Figure 1c). The PfEMP1 family divides into three main groups based on the receptor specificity of the N-terminal CIDR domain: (i) PfEMP1 proteins with CIDRα1 domains bind endothelial protein C receptor (EPCR), while (ii) PfEMP1 proteins with CIDRα2–6 domains bind CD36 and (iii) the atypical VAR2CSA PfEMP1 proteins bind placental chondroitin sulphate A (CSA) (Salanti *et al.*, 2004). In addition to these, a subset of PfEMP1 proteins 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 proteins with domain sub-variants NTSA-DBLα1-CIDRα1/β/γ/δ, whereas UPSB and UPSC *var* genes encode PfEMP1 proteins 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 outcomes of infection. Pregnancy-associated malaria is linked to parasites expressing VAR2CSA, whereas 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 proteins with unknown binding phenotypes of the N-terminal head structure and C-terminal PfEMP1 domains is largely unknown, albeit specific interactions with endothelial receptors and plasma proteins 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 a highly conserved *var1* variant of either type 3D7 or IT, which in most genomes occurs with a truncated or absent exon 2. Also, most genomes carry the unusually small and highly conserved *var3* genes, of unknown function (Figure 1c) (Otto *et al.*, 2019).

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. *Var* genes can be 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 provide a better link between *var* expression and PfEMP1 phenotype in *in vitro* assays, co-expression with other genes or gene families and epigenetics. While approaches for *var* assembly and quantification based on RNA-sequencing have recently been proposed (Wichers *et al.*, 2021; Stucke et al., 2021; Andrade et al., 2020; Tonkin-Hill *et al.*, 2018, Duffy et al., 2016), these 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 (Table 1).

Table 1: Comparison of previous *var* assembly approaches based on DNA- and RNA-sequencing.

Study	Assembler	k-mer	Transcript or gene assembly	Validation on reference strain(s) (Yes/No)	Validation on field strain(s) (Yes/No)	Validation across different expression levels (Yes/No)	Read length (Short/Long)	Read correction (Yes/No)	Scaffolding (Yes/No)	*Var* transcript filter approach	Assumption	Other limitations
Duffy *et al.*, 2016	Oases		Transcript	No	No	No	Short	Yes	Yes	Aligned to 399 *var* genes with BLAST (e-value < 10^{-5})		- No quantification of misassemblies - Unable to recover full length transcript assemblies
Dara *et al.*, 2017	Sprai and Celera (no longer maintained)	71	Gene	Yes (strain NF54)	Yes (12 UM patient samples)	NA – only genome assemblies	Long and short	Method assumes combination of long and short-read sequencing will identify errors	No	>500bp and aligned to VarDom database	Assumes a whole genome assembly is available	- Require prior filtering of human DNA - Need a combination of short-read and long-read sequencing
Tonkin-Hill *et al.*, 2018*	SoapDeNovo-Trans	21, 31, 41, 52 & 61	Transcript	Yes (strain ITG)	No	No	Short	No	No	>500bp and containing a sig. annotated *var* domain		- Unable to fully resolve N-terminus - No quantification of misassemblies
Otto *et al.*, 2019	Masurca + post-assembly improvements	Default	Gene	Yes (clone 3D7)	Yes (15 Pf3k reference genomes)	No	Short	Yes	Yes		Whole genome dataset	
Andrade *et al.*, 2020	Velvet	41	Transcript	No	No	No	Short	No	No	Aligned to VarDom database		- No quantification of misassemblies
Stucke *et al.*, 2021	rnaSPAdes	Default	Transcript	No	Yes (6 UM patient samples)	No – only the most expressed *var* gene	Short	Yes	Unclear	> 500bp and containing a sig. annotated *var* domain	Information about the true *var* annotation is available	- Performs *de novo* assembly on all non-human and *P. falciparum* mapping reads - Inconsistent results in 3 samples when comparing genomic and RNA-seq results for dominant *var* gene

* also used in Wichers *et al.*, 2021; Guillochon *et al.*, 2022
UM: uncomplicated malaria

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 the PfEMP1

129 phenotype of parasites isolated from malaria-infected donors (Pickford *et al.*, 2021, Joste *et al.*,
130 2020, Storm *et al.*, 2019, Tuikue Ndam *et al.*, 2017, Bruske *et al.*, 2016, Claessens *et al.*, 2012,
131 Lavstsen *et al.*, 2005, Jensen *et al.*, 2004, Kirchgatter & Portillo Hdel, 2002, Dimonte *et al.*, 2016,
132 Hoo *et al.*, 2019). However, *in vitro* conditions are considerably different to those found *in vivo,*
133 for example in terms of different nutrient availability and lack of a host immune response (Brown
134 & Guler, 2020). Previous studies found inconsistent results in terms of whether *var* gene
135 expression is impacted by culture and, if so, which *var* groups were the most affected (Zhang *et*
136 *al.*, 2011, Peters *et al.*, 2007). Similar challenges apply to the understanding of changes in *P.*
137 *falciparum* non-polymorphic core genes in culture, with the focus previously being on long-term
138 laboratory adapted parasites (Claessens *et al.*, 2017, Mackinnon *et al.*, 2009). Consequently, direct
139 interpretation of a short-term cultured parasite's transcriptome remains a challenge. It is
140 fundamental to understand *var* genes in context with the parasite's core transcriptome. This could
141 provide insights into *var* gene regulation and phenomena such as the proposed lower level of *var*
142 gene expression in asymptomatic individuals (Almelli *et al.*, 2014, Andrade *et al.*, 2020).
143
144 Here we present an improved method for assembly, characterization, and quantification of *var*
145 gene expression from RNA-sequencing data. This new approach overcomes previous limitations
146 and outperforms current methods, enabling a much greater understanding of the *var*
147 transcriptome. We demonstrate the power of this new approach by evaluating changes in *var*
148 gene expression of paired samples from clinical isolates *of P. falciparum* during their early
149 transition to *in vitro* culture*,* across several generations. The use of paired samples, which are
150 genetically identical and hence have the same *var* gene repertoire, allows validation of assembled
151 transcripts and direct comparisons of expression. We complement this with a comparison of
152 changes which occur in the non-polymorphic core transcriptome over the same transition into
153 culture. We find a background of modest changes in *var* gene expression with unpredictable
154 patterns of *var* gene switching, favouring an apparent convergence towards *var2csa* expression.
155 More extensive changes were observed in the core transcriptome during the first cycle of culture,
156 suggestive of a parasite stress response.
157
158 <u>Results</u>
159 To extend our ability to characterise *var* gene expression profiles and changes over time in clinical
160 *P. falciparum* isolates, we set out to improve current assembly methods. Previous methods for
161 assembling *var* transcripts have focussed on assembling whole transcripts (Tonkin-Hill *et al.*, 2018,
162 Wichers *et al.*, 2021, Guillochon *et al.*, 2022, Andrade *et al.*, 2020). However, due to the diversity
163 within PfEMP1 domains, their associations with disease severity and the fact different domain
164 types are not inherited together, a method focussing on domain assembly first was developed. In
165 addition, a novel whole transcript approach, using a different *de novo* assembler, was developed
166 and their performance compared to the method of Wichers *et al.* (hereafter termed "original
167 approach", Figure 2) (Wichers *et al.*, 2021). The new approaches made use of the MalariaGEN *P.*
168 *falciparum* dataset, which led to the identification of additional multi-mapping non-core reads (a
169 median of 3,955 reads per sample) prior to *var* transcript assembly (MalariaGen *et al.*, 2021). We
170 incorporated read error correction and improved large scaffold construction with fewer
171 misassemblies (see Methods). We then applied this pipeline to paired *ex vivo* and short-term *in*
172 *vitro* cultured parasites to enhance our understanding of the impact of short-term culturing on
173 the *var* transcriptome (Figure 3). The *var* transcriptome was assessed at several complementary
174 levels: first, changes in the dominantly expressed *var* gene and the homogeneity of the *var*
175 expression profile in paired samples were investigated; second, changes in *var* domain expression
176 through culture were assessed; and third, *var* group and global *var* gene expression changes were

evaluated. All these analyses on *var* expression were accompanied by analysis of the core transcriptome at the transition to short-term culture.

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 *P. falciparum* 3D7 time course RNA-sequencing dataset (European nucleotide archive (ENA): PRJEB31535) (Wichers *et al.*, 2019). The clinical dataset contained samples from 32 adult malaria patients, hospitalised in Hamburg, Germany (National Center for Biotechnology Information (NCBI) BioProject ID: PRJNA679547). Fifteen 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 *et al.*, 2021).

Our i) new whole transcript approach, ii) domain assembly approach, and iii) modified version of the original approach (see material and methods) were first applied to a *P. falciparum* 3D7 time course RNA-sequencing dataset to benchmark their performance (Wichers *et al.*, 2019) (Figure 2 – Figure supplement 1). The whole transcript approach performed best, achieving near perfect alignment scores for the dominantly expressed *var* gene (Figure 2 – Figure supplement 1a). 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 2 – Figure supplement 1c, f, g and h). However, we found high accuracies (> 0.95) across all approaches, meaning the sequences we assembled were correct (Figure 2 – Figure supplement 1b). The whole transcript approach also performed the best when assembling the lower expressed *var* genes (Figure 2 – Figure supplement 1e) and produced the fewest *var* chimeras compared to the original approach on *P. falciparum* 3D7. Fourteen misassemblies were observed with the whole transcript approach compared to 19 with the original approach (Table S1). This reduction in misassemblies was particularly apparent in the ring-stage samples.

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 2. The whole transcript approach produced the fewest transcripts, but of greater length than the domain approach and the original approach (Figure 2 – Figure supplement 2). The whole transcript approach also returned the largest N50 score (more than doubling the N50 of the original approach), which means that it was the most contiguous assembly produced. Remarkably, with the new whole transcript method, we observed a significant decrease (2 vs 336) in clearly misassembled transcripts with, for example, an N-terminal domain at an internal position.

Table 2: Statistics for the different approaches used to assemble the *var* transcripts. *Var* 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. Misassemblies represents the number of misassemblies for each approach. *Number of misassemblies were not determined for the domain approach due to its poor performance in other metrics.

	Number of contigs ≥500nts	Maximum length (nt)	Average contig length (nt)	N50	Number of misassemblies
Original approach	6,441	10,412	1,621	2,302	336
Domain approach	4,691	5,003	954	1,088	NA*
Whole transcript approach	3,011	12,586	2,771	5,381	2

When genome sequencing is not available, concordance of different *var* profiling approaches can support the validation of an approach. Here, the same methods used in the original analysis were applied for quantifying the expression of the assembled *var* transcripts and domains. This suggests any concordance in expression estimates likely reflects concordance at the domain annotation level. 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 2 – Figure supplement 3). As expected, comparable results were also seen for the differentially expressed transcripts identified in the original analysis between the naïve vs pre-exposed and severe vs non-severe comparisons, respectively (Figure 2 – Figure supplement 4).

Overall, the new whole transcript approach performed the best on the laboratory 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 clinical *ex vivo* samples (NCBI: PRJNA679547) (Wichers *et al.*, 2021). Therefore, it was selected for all subsequent analyses unless specified otherwise.

Establishing characterisation of *var* transcripts from *ex vivo* and *in vitro* samples

Of the 32 clinical isolates of *P. falciparum* from the German traveller dataset, 13 underwent one replication cycle of *in vitro* culture, 10 of these underwent a second generation and one underwent a third generation (Table 3). Most (9/13, 69%) isolates entering culture had a single MSP1 genotype, indicative of monoclonal infections. All samples were sequenced with a high read depth, although the *ex vivo* samples had a greater read depth than the *in vitro* samples (Table 3). Figure 3 shows a summary of the analysis performed.

Table 3: Summary of the clinical dataset used to analyze the impact of parasite culturing on gene expression. RNA-sequencing was performed on 32 malaria infected German traveler samples (Wichers *et al.*, 2021). The 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* cultivation 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.

			Generation		
		***Ex vivo* (n=32)**	**1 (n=13)**	**2 (n=10)**	**3 (n=1)**
Malaria exposure (n)	**Naïve**	15	6	4	1
	Previously exposed	17	7	6	0
Malaria severity (n)	**Severe**	8	3	1	0
	Non-severe	24	10	9	1
Number of MSP1 genotypes (number of samples)	**1**	22	9	7	1
	2	4	0	0	0
	3	5	3	0	0
	4	1	1	0	0
Number of *P. falciparum* PE* reads (non-*var*) (median, IQR) (million of reads)		34.6 (27.0–36.5)	17.1 (12.9–18.0)	17.2 (12.9–19.1)	15.1
Number of non-core *P. falciparum* PE* reads (median, IQR) (million of reads)		5.05 (3.62–6.60)	1.16 (1.07–1.40)	1.29 (1.04–1.58)	0.91
Number of assembled *var* contigs in a sample (≥500nts) (whole transcript approach) (median, IQR)		53 (44–84)	61 (38–76)	71.5 (48.25–79.5)	75
Number of assembled *var* contigs in a sample (≥1,500nts and 3 sig. domain annotations) (whole transcript approach) (median, IQR)		20 (7–31)	15.5 (10–26)	15 (10.25–23.75)	18

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* and *in vitro* samples (Figure 3 – Figure supplement 1a,b). 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 3 – Figure supplement 1c). The estimates produced from the mixture model approach showed high concordance with those observed by counting Giemsa-stained blood smears (Figure 3 – Figure supplement 1d). 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. 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 1500nt long and containing at least 3 significantly annotated *var* domains was the least restrictive, while the other approaches required the presence of a DBLα domain within the transcript. All three filtering approaches generated the same maximum length *var* transcript and similar N50 values. This suggests minimal differences in the three filtering approaches, whilst highlighting the importance of filtering assembled *var* transcripts.

In the original approach of Wichers *et al.*, 2021, the non-core reads of each sample used for *var* assembly were mapped against a pooled reference of assembled *var* transcripts from all samples, as a preliminary step towards differential *var* transcript expression analysis. This approach returned a small number of *var* transcripts which were expressed across multiple patient samples (Figure 3 – Figure supplement 2a). As genome sequencing was not available, it was not possible to know whether there was truly overlap in *var* genomic repertoires of the different patient samples, but substantial overlap was not expected. Stricter mapping approaches (for example, excluding transcripts shorter than 1500nt) changed the resulting *var* expression profiles and produced more realistic scenarios where similar *var* expression profiles were generated across paired samples, whilst there was decreasing overlap across different patient samples (Figure 3 – Figure supplement 2b,c). Given this limitation, we used the paired samples to analyse *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. The per patient approach showed consistent expression of *var* transcripts within samples from each patient but no overlap of *var* expression profiles across different patients (Figure 3 – Figure supplement 2d). Taken together, the per patient approach was better suited for assessing *var* transcriptional changes in longitudinal

samples. However, it has been hypothesised that more conserved *var* genes in field isolates increase parasite fitness during chronic infections, necessitating the need to correctly identify them (Dimonte *et al*., 2020, Otto *et al*., 2019). Accordingly, further work is needed to optimise the pooled sample approach to identify truly conserved *var* transcripts across different parasite isolates in cross-sectional studies.

Longitudinal analysis of *var* transcriptome from *ex vivo* to *in vitro* samples

To assess the changes in the *var* transcriptome induced by parasite culturing, we performed a series of analyses, all of which addressed different aspects: (i) changes in individual *var* gene expression pattern and *var* expression homogeneity ("per patient analysis"), (ii) changes in the expression of *var* variants conserved between strains, (iii) changes in the expression of PfEMP1 domains, (iv) changes in expression at the *var* group level, and (v) at the overall *var* expression level. We validated our results using the DBLα-tag approach and complemented the *var* -specific analysis by also examining changes in the core transcriptome.

To investigate whether dominant *var* gene expression changes through *in vitro* culture, rank analysis of *var* transcript expression was performed (Figure 4, Figure 4 – Figure supplement 1). In most cases a single dominant *var* transcript was detected. The dominant *var* gene did not change in most patient samples and the ranking of *var* gene expression remained similar. However, we observed a change in the dominant *var* gene being expressed through culture in isolates from three of 13 (23%) patients (#6, #17 and #26). Changes in the dominant *var* gene expression were also observed in the DBLα-tag data for these patients (described below). In parasites from three additional patients, #1, #7 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* sample. 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. Overall, the data suggest 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, *var* expression homogeneity (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 (patients #1, #2, #4) (Figure 4 – Figure supplement 2). 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 first-time infected severe malaria patients). Some of these samples (#1 and #26, both from first-time infected 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.

Expression of conserved *var* gene variants through short-term *in vitro* culture

Due to the relatively high level of conservation observed in *var1, var2csa* and *var3*, they do not present with the same limitations as regular *var* genes. Therefore, changes in their expression through short-term culture was investigated across all samples together. We observed no significant differences in the expression of conserved *var* gene variants, *var1-IT* (p_{adj} = 0.61, paired

356 Wilcoxon test), *var1-3D7* (p_{adj} =0.93, paired Wilcoxon test) and *var2csa* (p_{adj} =0.54, paired
357 Wilcoxon test) between paired *ex vivo* and generation 1 parasites, but *var2csa* was significantly
358 differentially expressed between generation 1 and generation 2 parasites (p_{adj} = 0.029, paired
359 Wilcoxon test) (Figure 4 – Figure supplement 3). However, *var2csa* expression previously
360 appeared to have decreased in some paired samples during the first cycle of cultivation (Figure 4
361 – Figure supplement 3).
362
363 Differential expression of *var* domains from *ex vivo* to *in vitro* samples
364 There is overlap in PfEMP1 domain subtypes of different parasite isolates which can be associated
365 with *var* gene groups and receptor binding phenotypes. This allows performing differential
366 expression analysis on the level of encoded PfEMP1 domain subtypes, as done in previous studies
367 (Tonkin-Hill *et al.*, 2018, Wichers *et al.*, 2021). PCA on *var* domain expression (Figure 5a) showed
368 some patients' *ex vivo* samples clustering away from their respective generation 1 sample (patient
369 #1, #2, #4, #12, #17, #25), again indicating a greater *var* transcriptional change relative to the
370 other samples during the first cycle of cultivation. However, in the pooled comparison of the
371 generation 1 vs *ex vivo* of all isolates, a single domain was significantly differentially expressed,
372 CIDRα2.5 associated with B-type PfEMP1 proteins and CD36-binding (Figure 5b). In the generation
373 2 vs *ex vivo* comparison, there were no domains significantly differentially expressed, however we
374 observed large $\log_2$FC values in similar domains to those changing most in the *ex vivo* vs generation
375 1 comparison (Figure 5c). No differentially expressed domains were found in the generation 1 vs
376 generation 2 comparison. These results suggest individual changes in *var* expression are not
377 reflected in the pooled analysis and the per patient approach is more suitable.
378
379 *Var* group expression analysis
380 A previous study found group A *var* genes to have a rapid transcriptional decline in culture
381 compared to group B *var* genes, however another study found a decrease in both group A and
382 group B *var* genes in culture (Zhang *et al.*, 2011, Peters *et al.*, 2007). These studies were limited as
383 the *var* type was determined by analysing the sequence diversity of DBLα domains, and by
384 quantitative PCR (qPCR) methodology which restricts analysis to quantification of
385 known/conserved sequences. Due to these results, the expression of group A *var* genes vs. group
386 B and C *var* genes was investigated using a paired analysis on all the DBLα (DBLα1 vs DBLα0 and
387 DBLα2) and NTS (NTSA vs NTSB) sequences assembled from *ex vivo* samples and across multiple
388 generations in culture. A linear model was created with group A expression as the response
389 variable, the generation and life cycle stage as independent variables and the patient information
390 included as a random effect. The same was performed using group B and C expression levels.
391
392 In both approaches, DBLα and NTS, we found no significant changes in total group A or group B
393 and C *var* gene expression levels (Figure 6). We observed high levels of group B and C *var* gene
394 expression compared to group A in all patients, both in the *ex vivo* samples and the *in vitro*
395 samples. In some patients we observed a decrease in group A *var* genes from *ex vivo* to generation
396 1 (patients #1, #2, #5, #6, #9, #12, #17, #26) (Figure 6a), however in all but four patients (patient
397 #1, #2, #5, #6) the levels of group B and C *var* genes remained consistently high from *ex vivo* to
398 generation 1 (Figure 6b). Interestingly, patients #6 and #17 also had a change in the dominant *var*
399 gene expression through culture. Taken together with the preceding results, it appears that
400 observed differences in *var* transcript expression occurring with transition to short-term culture
401 are not due to modulation of recognised *var* classes, but due to differences in expression of
402 particular *var* transcripts.
403
404 Quantification of total *var* gene expression

405 We observed a trend of decreasing total *var* gene expression between generations irrespective of
406 the assembler used in the analysis (Figure 6 – Figure supplement 1). A similar trend is seen with
407 the LARSFADIG count, which is commonly used as a proxy for the number of different *var* genes
408 expressed (Otto *et al.*, 2019). A linear model was created (using only paired samples from *ex vivo*
409 and generation 1) (Supplementary file 1) with proportion of total gene expression dedicated to
410 *var* gene expression as the response variable, the generation and life cycle stage as independent
411 variables and the patient information included as a random effect. This model showed no
412 significant differences between generations, suggesting that differences observed in the raw data
413 may be a consequence of small changes in developmental stage distribution in culture.
414
415 Validation of *var* expression profiling by DBLα-tag sequencing
416 Deep sequencing of RT-PCR-amplified DBLα expressed sequence tags (ESTs) combined with
417 prediction of the associated transcripts and their encoded domains using the Varia tool
418 (Mackenzie *et al.*, 2022) was performed to supplement the RNA-sequencing analysis. The raw
419 Varia output file is given in Supplementary file 2. Overall, we found a high agreement between the
420 detected DBLα-tag sequences and the *de novo* assembled *var* transcripts. A median of 96% (IQR:
421 93–100%) of all unique DBLα-tag sequences detected with >10 reads were found in the RNA-
422 sequencing approach. This is a significant improvement on the original approach (p= 0.0077,
423 paired Wilcoxon test), in which a median of 83% (IQR: 79–96%) was found (Wichers *et al.*, 2021).
424 To allow for a fair comparison of the >10 reads threshold used in the DBLα-tag approach, the
425 upper 75th percentile of the RNA-sequencing-assembled DBLα domains were analysed. A median
426 of 77.4% (IQR: 61–88%) of the upper 75th percentile of the assembled DBLα domains were found
427 in the DBLα-tag approach. This is a lower median percentage than the median of 81.3% (IQR: 73–
428 98%) found in the original analysis (p= 0.28, paired Wilcoxon test) and suggests the new assembly
429 approach is better at capturing all expressed DBLα domains.
430
431 The new whole transcript assembly approach also had high consistency with the domain
432 annotations predicted from Varia. Varia predicts *var* sequences and domain annotations based on
433 short sequence tags, using a database of previously defined *var* sequences and annotations
434 (Mackenzie *et al.*, 2022). A median of 85% of the DBLα annotations and 73% of the DBLα-CIDR
435 domain annotations, respectively, identified using the DBLα-tag approach were found in the RNA
436 sequencing approach. This further confirms the performance of the whole transcript approach
437 and it was not restricted by the pooled approach of the original analysis. We also observed
438 consistent results with the per patient analysis, in terms of changes in the dominant *var* gene
439 expression (described above) (Supplementary file 2). In line with the RNA-sequencing data, the
440 DBLα-tag approach revealed no significant differences in Group A and Group B and C groups during
441 short-term culture, further highlighting the agreement of both methods (Figure 6 – Figure
442 supplement 2).
443
444 Differential expression analysis of the core transcriptome between *ex vivo* and *in vitro* samples
445 Given the modest changes in *var* gene expression repertoire upon culture we wanted to
446 investigate the extent of any accompanying changes in the core parasite transcriptome. PCA was
447 performed on core gene (*var, rif, stevor, surf* and rRNA genes removed) expression, adjusted for
448 life cycle stage. We observed distinct clustering of *ex vivo*, generation 1, and generation 2 samples,
449 with patient identity having much less influence (Figure 7a). There was also a change from the
450 heterogeneity between the *ex vivo* samples to more uniform clustering of the generation 1
451 samples (Figure 7a), suggesting that during the first cycle of cultivation the core transcriptomes of
452 different parasite isolates become more alike.
453

454 In total, 920 core genes (19% of the core transcriptome) were found to be differentially expressed
455 after adjusting for life cycle stages using the mixture model approach between *ex vivo and*
456 generation 1 samples (Supplementary file 3). The majority were upregulated, indicating a
457 substantial transcriptional change during the first cycle of *in vitro* cultivation (Figure 7b). 74 genes
458 were found to be upregulated in generation 2 when compared to the *ex vivo* samples, many with
459 $\log_2$FC greater than those in the *ex vivo* vs generation 1 comparison (Figure 7c). No genes were
460 found to be significantly differentially expressed between generation 1 and generation 2.
461 However, five genes had a $\log_2$FC >= 2 and were all upregulated in generation 2 compared to
462 generation 1. Interestingly, the gene with the greatest fold change, encoding ROM3
463 (PF3D7_0828000), was also found to be significantly downregulated in generation 1 parasites in
464 the *ex vivo* vs generation 1 analysis. The other four genes were also found to be non-significantly
465 downregulated in generation 1 parasites in the *ex vivo* vs generation 1 analysis. This suggests
466 changes in gene expression during the first cycle of cultivation are the greatest compared to the
467 other cycles.
468
469 The most significantly upregulated genes (in terms of fold change) in generation 1 contained
470 several small nuclear RNAs, splicesomal RNAs and non-coding RNAs (ncRNAs). 16 ncRNAs were
471 found upregulated in generation 1, with several RNA-associated proteins having large fold changes
472 ($\log_2$FC > 7) Significant gene ontology (GO) terms and Kyoto encyclopedia of genes and genomes
473 (KEGG) pathways for the core genes upregulated in generation 1 included "entry into host",
474 "movement into host" and "cytoskeletal organisation" suggesting the parasites undergo a change
475 in invasion efficiency, which is connected to the cytoskeleton, during their first cycle of *in vitro*
476 cultivation (Figure 7 – Figure supplement 1). We observed eight AP2 transcription factors
477 upregulated in generation 1 (PF3D7_0404100/AP2-SP2, PF3D7_0604100/SIP2,
478 PF3D7_0611200/AP2-EXP2, PF3D7_0613800, PF3D7_0802100/AP2-LT, PF3D7_1143100/AP2-O,
479 PF3D7_1239200, PF3D7_1456000/AP2-HC) with no AP2 transcription factors found to be
480 downregulated in generation 1. To confirm the core gene expression changes identified were not
481 due to the increase in parasite age during culture, as indicated by upregulation of many schizont-
482 related genes, core gene differential expression analysis was performed on paired *ex vivo* and
483 generation 1 samples that contained no schizonts or gametocytes in generation 1. The same genes
484 were identified as significantly differentially expressed with a Spearman's rank correlation of 0.99
485 for the $\log_2$FC correlation between this restricted sample approach and those produced using all
486 samples (Figure 7 – Figure supplement 2).
487
488 <u>Cultured parasites as surrogates for assessing the *in vivo* core gene transcriptome</u>
489 In the original analysis of *ex vivo samples*, hundreds of core genes were identified as significantly
490 differentially expressed between pre-exposed and naïve malaria patients. We investigated
491 whether these differences persisted after *in vitro* cultivation. We performed differential
492 expression analysis comparing parasite isolates from naïve (n=6) vs pre-exposed (n=7) patients,
493 first between their *ex vivo* samples, and then between the corresponding generation 1 samples.
494 Interestingly, when using the *ex vivo* samples, we observed 206 core genes significantly
495 upregulated in naïve patients compared to pre-exposed patients (Figure 7 – Figure supplement
496 3a). Conversely, we observed no differentially expressed genes in the naïve vs pre-exposed
497 analysis of the paired generation 1 samples (Figure 7 – Figure supplement 3b). Taken together
498 with the preceding findings, this suggests one cycle of cultivation shifts the core transcriptomes
499 of parasites to be more alike each other, diminishing inferences about parasite biology *in vivo*.
500
501 A summary describing the rationale, results and interpretation of each approach in our analysis
502 pipeline can be found in Table 4.

503 **Table 4: Summary of the different levels of analysis performed to assess the effect of short-term parasite culturing on *var***
504 **and core gene expression, their rational, method, results, and interpretation.**

Analysis level	Analysis	Rationale	Method	Results	Interpretation
***var* transcript**	Per patient expression ranking	Relative quantification of *var* transcripts over consecutive generations of parasites originating from the same patient to reveal *var* gene switching events	Combine assembled *var* transcripts for each patient into a reference and quantify expression, validated with DBLα-tag analysis	46% of the patient samples had a change in the dominant or top 3 highest expressed *var* gene	Modest changes in most samples, but unpredictable *var* gene switching during culture in some samples
	Per patient *var* expression homogeneity (VEH)	Determine the overall diversity of *var* gene expression (number of different variants expressed and their abundance) to assess impact of culturing on the overall *var* gene expression pattern	Comparison of diversity curves based on per patient quantification of the *var* transcriptome	39% of *ex vivo* samples diversity curves distinct from *in vitro* samples	Some patient samples underwent a much greater *var* transcriptional change compared to others
	Conserved *var* variants	Assessing and comparing the expression levels of strain-transcendent *var* gene variants (*var1*, *var2csa*, *var3*) between samples	Reference of all assembled transcripts for each conserved *var* gene and quantify expression	*var2csa* expression increases in 2nd *in vitro* generation	Parasites converge to *var2csa* during short-term *in vitro* culture
***var*-encoded PfEMP1 domains**	Differential expression of PfEMP1 domains	Identification, quantification and comparison of expression levels of different *var* gene-encoded PfEMP1 domains associated with different disease manifestations	Pool all assembled *var* transcripts into a reference and quantify expression of each domain	46% of the *ex vivo* samples cluster away from their *in vitro* samples in PCA plots, distinct clustering by *in vitro* generation was not observed; CIDRα2.5 significantly differentially expressed between *ex vivo* and generation 1	Transition to culture results in modest modulation of particular *var* domains
***var* group**	Expression of NTS (NTSA vs NTSB) and DBLα (DBLα1 vs DBLα0+ DBLα2)	Quantification and comparison of expression levels of different *var* gene groups (group A vs. group B and C)	Create a reference of all assembled DBLα and NTS domains for each patient and quantify expression. Validated with DBLα-tag analysis	No significant changes	No preferential up or down regulation of certain *var* groups during transition to culture
Global *var* expression	LARSFADIG coverage	Assessing the overall *var* gene expression level (excluding *var2csa*)	Assemble the LARSFADIG motif and map non-core reads to quantify coverage	Trend for decrease in global *var* expression during culture, but no significant changes	Subtle reduction in global *var* gene expression may reflect increase in parasite age during culture
Core genes	Differential gene expression (DGE)	Assessing the impact of cultivation on the parasite core gene transcriptome	Differential expression analysis of core genes (*P. falciparum* 3D7 used as reference)	19% of the core transcriptome significantly differentially expressed between paired *ex vivo* and generation 1 *in vitro* samples; distinct clustering by parasite generation observed; upregulation of invasion and replication related genes *in vitro*	Parasites core gene expression changes substantially upon entering culture

505

506 ## Discussion

507 Multiple lines of evidence point to PfEMP1 as a major determinant of malaria pathogenesis, but
508 previous approaches for characterising *var* expression profiles in field samples have limited *in vivo*
509 studies of PfEMP1 function, regulation, and association with clinical symptoms (Tarr *et al.*, 2018,
510 Lee *et al.*, 2018, Warimwe *et al.*, 2013, Rorick *et al.*, 2013, Zhang *et al.*, 2011, Taylor *et al.*, 2002).
511 A more recent approach, based on RNA-sequencing, overcame many of the limitations imposed
512 by the previous primer-based methods (Tonkin-Hill *et al.*, 2018, Wichers *et al.*, 2021). However,
513 depending on the expression level and sequencing depth, *var* transcripts were found to be
514 fragmented and only a partial reconstruction of the *var* transcriptome was achieved (Tonkin-Hill
515 *et al.*, 2018, Wichers *et al.*, 2021, Andrade *et al.*, 2020, Guillochon *et al.*, 2022, Yamagishi *et al.*,
516 2014). The present study developed a novel approach for *var* gene assembly and quantification
517 that overcomes many of these limitations.
518
519 Our new approach used the most geographically diverse reference of *var* gene sequences to date,
520 which improved the identification of reads derived from *var* transcripts. This is crucial when
521 analysing patient samples with low parasitaemia where *var* transcripts are hard to assemble due
522 to their low abundancy (Guillochon *et al.*, 2022). Our approach has wide utility due to stable
523 performance on both laboratory-adapted and clinical samples. Concordance in the different *var*
524 expression profiling approaches (RNA-sequencing and DBLα-tag) on *ex vivo* samples increased
525 using the new approach by 13%, when compared to the original approach (96% in the whole
526 transcript approach compared to 83% in Wichers *et al.*, 2021. This suggests the new approach
527 provides a more accurate method for characterising *var* genes, especially in samples collected
528 directly from patients. Ultimately, this will allow a deeper understanding of relationships between
529 *var* gene expression and clinical manifestations of malaria.
530
531 Having a low number of long contigs is desirable in any *de novo* assembly. This reflects a
532 continuous assembly, as opposed to a highly fragmented one where polymorphic and repeat
533 regions could not be resolved (Lischer & Shimizu, 2017). An excessive number of contigs cannot
534 be reasonably handled computationally and results from a high level of ambiguity in the assembly
535 (Yang *et al.*, 2012). We observed a greater than 50% reduction in the number of contigs produced
536 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 of 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 PfEMP1 proteins.

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 can be applied for example in 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.

This study had substantial power through the use of paired samples. However, many *var* gene expression studies do not have longitudinal sampling. Future work should focus on identifying the best approach for analysing the *var* transcripts in cross-sectional samples. Higher level *var* classification systems, such as the PfEMP1 predicted binding phenotype or domain cassettes, could be applied to test for over-representation of different *var* gene features in different groups of interest, because the assumption of overlapping *var* repertoires at these levels of classification would be more realistic. This was briefly explored in our analysis through *var* domain differential expression analysis, which found minimal changes in *var* domain expression through short-term culture, supporting the per patient analysis results. This could be further improved by advancing the classifications of domain subtypes. This has recently been studied using MEME to identify short nucleotide motifs that are representative of domain subtypes (Otto et al., 2019). Other research could investigate clustering *var* transcripts based on sequence identity and testing for clusters associated with specific malaria disease groups.

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. However, the conserved *var2csa* 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 is suggested to be the default *var* gene 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

586 results are consistent with the role of *var2csa* as the dominant "sink node" (Zhang *et al.*, 2022,
587 Ukaegbu *et al.*, 2015, Ukaegbu *et al.*, 2014, Mok *et al.*, 2008). A previous study suggested *in vitro*
588 cultivation of controlled human malaria infection samples resulted in dramatic changes in *var* gene
589 expression (Lavstsen *et al.*, 2005, Peters *et al.*, 2007). Almost a quarter of samples in our analysis
590 showed more pronounced and unpredictable changes. In these individuals, the dominant *var* gene
591 being expressed changed within one cycle of cultivation. This implies short-term culture can result
592 in unpredictable *var* gene expression as observed previously using a semi-quantitative RT-PCR
593 approach (Bachmann *et al.*, 2011) and that one would need to confirm *in vivo* expression matches
594 *in vitro* expression. This can be achieved using the assembly approach described here.
595
596 We observed no generalised pattern of up- or downregulation of specific *var* groups following
597 transition to culture. This implies there is probably not a selection event occurring during culture
598 but may represent a loss of selection that is present *in vivo.* A global down regulation of certain
599 *var* groups might only occur as a selective process over many cycles in extended culture.
600 Determining changes in *var* group expression levels are difficult using degenerate qPCR primers
601 bias and previous studies have found conflicting results in terms of changes of expression of *var*
602 groups through cultivation. Zhang *et al.*, 2011 found a rapid transcriptional decline of group A and
603 group B *var* genes, however Peters *et al.*, 2007 found group A *var* genes to have a high rate of
604 downregulation, when compared to group B *var* genes. These studies differed in the stage
605 distribution of the parasites and were limited in measuring enough variants through their use of
606 primers. Our new approach allowed for the identification of more sequences, with 26.6% of
607 assembled DBLα domains not found via the DBLα-tag approach. This better coverage of the
608 expressed *var* diversity was not possible in these previous studies and may explain discrepancies
609 observed.
610
611 Generally, there was a high consensus between all levels of *var* gene analysis and changes
612 observed during short-term *in vitro* cultivation. However, the impact of short-term culture was
613 the most apparent at the *var* transcript level and became less clear at the *var* domain, *var* type
614 and global *var* gene expression level. This highlights the need for accurate characterisation of full
615 length *var* transcripts and analysis of the *var* transcriptome at different levels, both of which can
616 be achieved with the new approach developed here.
617
618 We saw striking changes in the core gene transcriptomes between *ex vivo* and generation 1
619 parasites with 19% of the core genome being differentially expressed. A previous study showed
620 that expression of 18% of core genes were significantly altered after ~50 cycles through culture
621 (Mackinnon *et al.*, 2009), but our data suggest that much of this change occurs early in the
622 transition to culture. We observed genes with functions unrelated to ring-stage parasites were
623 among those most significantly expressed in the generation 1 vs *ex vivo* analysis, suggesting the
624 culture conditions may temporarily dysregulate stage-specific expression patterns or result in the
625 parasites undergoing a rapid adaptation response (Andreadaki *et al.*, 2020, Beeson *et al.*, 2016).
626 Several AP2 transcription factors (AP2-SP2, AP2-EXP2, AP2-LT, AP2-O and AP2-HC) were
627 upregulated in generation 1. AP2-HC has been shown to be expressed in asexual parasites
628 (Carrington *et al.*, 2021). AP2-O is thought to be specific for the ookinete stage and AP2-SP2 plays
629 a key role in sporozoite stage specific gene expression (Kaneko *et al.*, 2015, Yuda *et al.*, 2010). Our
630 findings are consistent with another study investigating the impact of long-term culture
631 (Mackinnon *et al.*, 2009) which also found genes like merozoite surface proteins differentially
632 expressed, however they were downregulated in long-term cultured parasites, whereas we found
633 them upregulated in generation 1. This suggests short-term cultured parasites might be
634 transcriptionally different from long-term cultured parasites, especially in their invasion

635 capabilities, something previously unobserved. Several genes involved in the stress response of
636 parasites were upregulated in generation 1, for example DnaJ proteins, serine proteases and ATP
637 dependent CLP proteases (Oakley *et al.*, 2007). The similarity of the core transcriptomes of the *in*
638 *vitro* samples compared to the heterogeneity seen in the *ex vivo* samples could be explained by a
639 stress response upon entry to culture. Studies investigating whether the dysregulation of stage
640 specific expression and the expression of stress associated genes persist in long-term culture are
641 required to understand whether they are important for growth in culture. Critically, the marked
642 differences presented here suggest the impact of short-term culture can override differences
643 observed in both the *in vivo* core and *var* transcriptomes of different disease manifestations.
644
645 In summary, we present an enhanced approach for *var* transcript assembly which allows for *var*
646 gene expression to be studied in connection to *P. falciparum's* core transcriptome through RNA-
647 sequencing. This will be useful for expanding our understanding of *var* gene regulation and
648 function in *in vivo* samples. As an example of the capabilities of the new approach, the method
649 was used to quantify differences in gene expression upon short-term culture adaptation. This
650 revealed that inferences from clinical isolates of *P. falciparum* put into short-term culture must be
651 made with a degree of caution. Whilst *var* gene expression is often maintained, unpredictable
652 switching does occur, necessitating that the similarity of *in vivo* and *in vitro* expression should be
653 confirmed. The more extreme changes in the core transcriptome could have much bigger
654 implications for understanding other aspects of parasite biology such as growth rates and drug
655 susceptibility and raise a need for additional caution. Further work is needed to examine *var* and
656 core transcriptome changes during longer term culture on a larger sample size. Understanding the
657 ground truth of the *var* expression repertoire of *Plasmodium* field isolates still presents a unique
658 challenge and this work expands the database of *var* sequences globally. The increase in long-read
659 sequencing and the growing size of *var* gene databases containing isolates from across the globe
660 will help overcome this issue in future studies.
661
662 ## Materials and Methods
663
664 **Table 5: Key Resources Table.**
665

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Plasmodium falciparum	Isolate #1	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #2	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #3	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #4	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #5	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #6	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #7	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #9	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #10	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #11	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #12	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #13	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #14	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #15	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe

Plasmodium falciparum	Isolate #16	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #17	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #18	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #19	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #20	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #21	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #22	Patient isolate (Wichers *et al.*, 2021)		Naïve, non-severe
Plasmodium falciparum	Isolate #23	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #24	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #25	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #26	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #27	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #28	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #29	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #30	Patient isolate (Wichers *et al.*, 2021)		Naïve, severe
Plasmodium falciparum	Isolate #31	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #32	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
Plasmodium falciparum	Isolate #33	Patient isolate (Wichers *et al.*, 2021)		Pre-exposed, non-severe
others	Plasmodipur filter	EuroProxima	Transia Cat. #: 8011Filter25u	
commercial reagent	TRIzol	ThermoFischer	Cat. #: 15596026	
commercial kit	RNeasy MinElute Cleanup Kit	Qiagen	Cat. #: 74204	
commercial kit	GLOBINclear kit	ThermoFischer	Cat. #: AM1980	
commercial kit	NEBNext Poly(A) mRNA Magnetic Isolation Module	NEB	Cat. #: E7490L	
commercial kit	NEBNext Ultra RNA Library Prep Kit for Illumina	NEB	Cat. #: E7530L	
commercial kit	NEBNext Multiplex Oligos for Illumina (Index Primers Set 1)	NEB	Cat. #: E7335L	
commercial kit	KAPA HiFi plus dNTPs	Roche	Cat. #: 7958846001	
software, algorithm	SoapDeNovo-Trans	SoapDeNovo-Trans	RRID:SCR_013268	
software, algorithm	rnaSPAdes	rnaSPAdes	RRID:SCR_016992	
software, algorithm	Oases	Oases	RRID:SCR_011896	
software, algorithm	DESeq2	DESeq2	RRID:SCR_015687	
software, algorithm	Salmon	Salmon	RRID:SCR_017036	
software, algorithm	CAP3	CAP3	RRID:SCR_007250	
software, algorithm	SSPACE	SSPACE	RRID:SCR_005056	
software, algorithm	featureCounts	featureCounts	RRID:SCR_012919	

666

667 Materials availability statement

668 The authors confirm that the data supporting the findings of this study are available within the
669 article and its supplementary materials. Additionally, the raw sequencing data are openly available
670 in National Center for Biotechnology Information (NCBI) at BioProject ID: PRJNA679547
671 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA679547). The already published

672 laboratory dataset (3D7 time course RNA-sequencing data) was deposited in the European
673 nucleotide archive (ENA): PRJEB31535 (https://www.ebi.ac.uk/ena/browser/view/PRJEB31535)
674 (Wichers *et al.*, 2019) and is also available on plasmoDB. All scripts are available in the GitHub
675 repository (https://github.com/ClareAndradiBrown/varAssembly).
676
677 Ethics statement
678 The study was conducted according to the principles of the Declaration of Helsinki, 6th edition,
679 and the International Conference on Harmonization-Good Clinical Practice (ICH-GCP) guidelines.
680 All 32 patients were treated as inpatients or outpatients in Hamburg, Germany (outpatient clinic
681 of the University Medical Center Hamburg-Eppendorf (UKE) at the Bernhard Nocht Institute for
682 Tropical Medicine, UKE, Bundeswehrkrankenhaus) (Wichers *et al.*, 2021). Blood samples for this
683 analysis were collected after patients had been informed about the aims and risks of the study
684 and had signed an informed consent form for voluntary blood collection (n=21). In the remaining
685 cases, no intended blood samples were collected but residuals from diagnostic blood samples
686 were used (n=11). The study was approved by the responsible ethics committee (Ethics
687 Committee of the Hamburg Medical Association, reference numbers PV3828 and PV4539).
688
689 Blood sampling, processing and *in vitro* cultivation of *P. falciparum*
690 EDTA blood samples (1–30 mL) were collected from 32 adult falciparum malaria patients for *ex*
691 *vivo* transcriptome profiling as reported by Wichers *et al.*, 2021 (Wichers *et al.*, 2021), hereafter
692 termed "the original analysis". Blood was drawn and either immediately processed (#1, #2, #3, #4,
693 #11, #12, #14, #17, #21, #23, #28, #29, #30, #31, #32) or stored overnight at 4°C until processing
694 (#5, #6, #7, #9, #10, #13, #15, #16, #18, #19, #20, #22, #24, #25, #26, #27, #33). If samples were
695 stored overnight, the *ex vivo* and *in vitro* samples were still processed at the same time (so paired
696 samples had similar storage). Erythrocytes were isolated by Ficoll gradient centrifugation, followed
697 by filtration through Plasmodipur filters (EuroProxima) to remove residual granulocytes. At least
698 400 µl of the purified erythrocytes were quickly lysed in 5 volumes of pre-warmed TRIzol
699 (ThermoFisher Scientific) and stored at -80°C until further processing ("*ex vivo* samples"). When
700 available, the remainder was then transferred to *in vitro* culture either without the addition of
701 allogeneic red cells or with the addition of O+ human red cells (blood bank, UKE) for dilution
702 according to a protocol adopted from Trager and Jensen (Table S4). Cultures were maintained at
703 37°C in an atmosphere of 1% O2, 5% CO2, and 94% N2 using RPMI complete medium containing
704 10% heat-inactivated human serum (A+, Interstate Blood Bank, Inc., Memphis, USA). Cultures
705 were sampled for RNA purification at the ring stage by microscopic observation of the individual
706 growth of parasite isolates, and harvesting was performed at the appropriate time without prior
707 synchronization treatment ("*in vitro* samples"). 13 of these *ex vivo* samples underwent one cycle
708 of *in vitro* cultivation, ten of these generation 1 samples underwent a second cycle of *in vitro*
709 cultivation. One of these generation 2 samples underwent a third cycle of *in vitro* cultivation (Table
710 2). In addition, an aliquot of *ex vivo* erythrocytes (approximately 50–100 µl) and aliquots of *in vitro*
711 cell cultures collected as indicated in Supplementary file 4 were processed for gDNA purification
712 and MSP1 genotyping as described elsewhere (Wichers *et al.*, 2021, Robert *et al.*, 1996).
713
714 RNA purification, RNA-sequencing library preparation, and sequencing
715 RNA purification was performed as described in Wichers *et al.*, 2021, using TRIzol in combination
716 with the RNeasy MinElute Kit (Qiagen) and DNase digestion (DNase I, Qiagen). Human globin
717 mRNA was depleted from all samples except from samples #1 and #2 using the GLOBINclear kit
718 (ThermoFisher Scientific). The median RIN value over all *ex vivo* samples was 6.75 (IQR: 5.93–7.40),
719 although this measurement has only limited significance for samples containing RNA of two
720 species. Accordingly, the RIN value increased upon cultivation for all *in vitro* samples

721 (Supplementary file 5). Customized library construction in accordance to Tonkin-Hill *et al.,* 2018,
722 including amplification with KAPA polymerase and HiSeq 2500 125 bp paired-end sequencing was
723 performed by BGI Genomics Co. (Hong Kong).
724
725 Methods for assembling *var* genes
726 Previously Oases, Velvet, SoapDeNovo-Trans or MaSuRCA have been used for *var* transcript
727 assembly (Wichers *et al.*, 2021, Andrade *et al.*, 2020, Otto *et al.*, 2019, Tonkin-Hill *et al.*, 2018).
728 Previous methods either did not incorporate read error correction or focussed on gene assembly,
729 as opposed to transcript assembly (Schulz *et al.*, 2012, Zerbino & Birney, 2008, Xie *et al.*, 2014,
730 Zimin *et al.*, 2013). Read error correction is important for *var* transcript assembly due to the highly
731 repetitive nature of the *P. falciparum* genome. Recent methods have also focused on whole
732 transcript assembly, as opposed to initial separate domain assembly followed by transcript
733 assembly (Wichers *et al.*, 2021, Andrade *et al.*, 2020, Otto *et al.*, 2019, Tonkin-Hill *et al.*, 2018).
734 The original analysis used SoapDeNovo-Trans to assemble the *var* transcripts, however it is
735 currently not possible to run all steps in the original approach, due to certain tools being improved
736 and updated. Therefore, SoapDeNovo-Trans (k=71) was used and termed the original approach.
737
738 Here, two novel methods for whole *var* transcript and *var* domain assembly were developed and
739 their performance was evaluated in comparison to the original approach (Figure 2b). In both
740 methods the reads were first mapped to the human g38 genome and any mapped reads were
741 removed. Next, the unmapped reads were mapped to a modified *P. falciparum* 3D7 genome with
742 *var* genes removed, to identify multi-mapping reads commonly present in *Plasmodium* RNA-
743 sequencing datasets. Any mapped reads were removed. In parallel, the unmapped RNA reads from
744 the human mapping stage were mapped against a reference of field isolate *var* exon 1 sequences
745 and the mapped reads identified (Otto *et al.*, 2019). These reads were combined with the
746 unmapped reads from the 3D7 genome mapping stage and duplicate reads removed. All mapping
747 was performed using sub-read align as in the original analysis (Wichers *et al.*, 2021). The reads
748 identified at the end of this process are referred to as "non-core reads".
749
750 Whole *var* transcript and *var* domain assembly methods
751 For whole *var* transcript assembly the non-core reads, for each sample separately, were
752 assembled using rnaSPAdes (k-mer =71, read_error_correction on) (Bushmanova *et al.*, 2019).
753 Contigs were joined into larger scaffolds using SSPACE (parameters -n 31 -x 0 -k 10) (Boetzer *et al.*,
754 2011). Transcripts < 500nt were excluded, as in the original approach. The included transcripts
755 were annotated using hidden Markov models (HMM) (Finn *et al.*, 2011) built on the *Rask et al.*,
756 2010 dataset and used in Tonkin-Hill *et al.*, 2018. When annotating the whole transcript, the most
757 significant alignment was taken as the best annotation for each region of the assembled transcript
758 (e-value cut off 1e-5). Multiple annotations were allowed on the transcript if they were not
759 overlapping, determined using cath-resolve-hits (Lewis *et al.*, 2019). Scripts are available in the
760 GitHub repository (https://github.com/ClareAndradiBrown/varAssembly)
761
762 In the *var* domain assembly approach, separate domains were assembled first and then joined up
763 to form transcripts. First, the non-core reads were mapped (nucleotide basic local alignment tool
764 (blastn) short read option) to the domain sequences as defined in Rask *et al.*, 2010. This was found
765 to produce similar results when compared to using tblastx. An e-value threshold of 1e-30 was used
766 for the more conserved DBLα domains and an e-value of 1e-10 for the other domains. Next, the
767 reads mapping to the different domains were assembled separately. rnaSPAdes
768 (read_error_correction on, k-mer = 15), Oases (kmer = 15) and SoapDeNovo2 (kmer = 15) were all
769 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 $\sqrt{\text{accuracy}}$* 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 same method that was used in the original analysis was applied for quantifying the expression of the assembled *var* transcripts. 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 in the original analysis (Love *et al.*, 2014). Here, the same approach, as used in the original analysis, was applied to see if concordant expression estimates were obtained. As genomic sequencing was not available, this provided a confirmation of the whole transcript approach after the domain annotation step. The assembled *var* transcripts produced by the whole

819 transcript assembly approach had their expression quantified at the transcript and domain level,
820 as in the original method, and the results were compared to those obtained by the original
821 method. To quantify domain expression, featureCounts was used, as in the original method with
822 the counts for each domain aggregated (Liao *et al.*, 2014). Correlation analysis between the
823 domain's counts from the whole transcript approach and the original method was performed for
824 each *ex vivo* sample. Differential expression analysis was also performed using DESeq2, as in the
825 original analysis and the results compared (Love *et al.*, 2014, Wichers *et al.*, 2021).
826
827 Estimation of parasite lifecycle stage distribution in *ex vivo* and short-term *in vitro* samples
828 To determine the parasite life cycle stage proportions for each sample the mixture model
829 approach of the original analysis (Tonkin-Hill *et al.*, 2018, Wichers *et al.*, 2021) and the SCDC
830 approach were used (Dong *et al.*, 2021, Howick *et al.*, 2019). Recently, it has been determined
831 that species-agnostic reference datasets can be used for efficient and accurate gene expression
832 deconvolution of bulk RNA-sequencing data from any *Plasmodium* species and for correct gene
833 expression analyses for biases caused by differences in stage composition among samples (Tebben
834 *et al.*, 2022). Therefore, the *Plasmodium berghei* single cell atlas was used as reference with
835 restriction to 1:1 orthologs between *P. berghei* and *P. falciparum*. This reference was chosen as it
836 contained reference transcriptomes for the gametocyte stage. To ensure consistency with the
837 original analysis, proportions from the mixture model approach were used for all subsequent
838 analyses (Wichers *et al.*, 2021). For comparison, the proportion of different stages of the parasite
839 life cycle in the *ex vivo* and *in vitro* samples was determined by two independent readers in
840 Giemsa-stained thin blood smears. The same classification as the mixture model approach was
841 used (8, 19, 30, and 42 hours post infection corresponding to ring, early trophozoite, late
842 trophozoite and schizont stages respectively). Significant differences in ring stage proportions
843 were tested using pairwise Wilcoxon tests. For the other stages, a modified Wilcoxon rank test for
844 zero-inflated data was used (Wang *et al.*, 2021). *Var* gene expression is highly stage dependent,
845 so any quantitative comparison between samples needs adjustment for developmental stage. The
846 life cycle stage proportions determined from the mixture model approach were used for
847 adjustment.
848
849 Characterising *var* transcripts
850 The whole transcript approach was applied to the paired *ex vivo* and *in vitro* samples. Significant
851 differences in the number of assembled *var* transcripts and the length of the transcripts across
852 the generations was tested using the paired Wilcoxon test. Redundancy was removed from the
853 assembled *var* transcripts and transcripts and domains were quantified using the approach
854 described above. Three additional filtering steps were applied separately to this reference of
855 assembled *var* transcripts to ensure the *var* transcripts that went on to have their expression
856 quantified represented true *var* transcripts. The first method restricted *var* transcripts to those
857 greater than 1500nt containing at least 3 significantly annotated *var* domains, one of which had
858 to be a DBLα domain. The second restricted *var* transcripts to those greater than 1500nt and
859 containing a DBLα domain. The third approach restricted *var* transcripts to those greater than
860 1500nt with at least 3 significant *var* domain annotations.
861
862 Per patient *var* transcript expression
863 A limitation of *var* transcript differential expression analysis is that it assumes all *var* sequences
864 have the possibility of being expressed in all samples. However, since each parasite isolate has a
865 different set of *var* gene sequences, this assumption is not completely valid. To account for this,
866 *var* transcript expression analysis was performed on a per patient basis. For each patient, the
867 paired *ex vivo* and *in vitro* samples were analysed. The assembled *var* transcripts (at least 1500nt

868 and containing 3 significantly annotated *var* domains) across all the generations for a patient were
869 combined into a reference, redundancy was removed as described above, and expression was
870 quantified using Salmon (Patro *et al.*, 2017). *Var* transcript expression was ranked, and the
871 rankings compared across the generations.
872
873 *Var* expression homogeneity (VEH)
874 VEH is defined as the extent to which a small number of *var* gene sequences dominate an isolate's
875 expression profile (Warimwe *et al.*, 2013). Previously, this has been evaluated by calculating a
876 commonly used α diversity index, the Simpson's index of diversity. Different α diversity indexes
877 put different weights on evenness and richness. To overcome the issue of choosing one metric, α
878 diversity curves were calculated (Wagner *et al.*, 2018). Equation 1 is the computational formula
879 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
880 analysis represented the proportion of *var* gene expression dedicated to *var* transcript *k*. *q*
881 determined how much weight is given to rare vs abundant *var* transcripts. The smaller the *q* value,
882 the less weight was given to the more abundant *var* transcript. VEH was investigated on a per
883 patient basis.
884
885
886 Equation 1 $$D_{(q)} = (\sum_{k=1}^{K} p_k^q)^{\frac{1}{1-q}}$$
887
888
889 Conserved *var* gene variants
890 To check for the differential expression of conserved *var* gene variants *var1-3D7*, *var1-IT* and
891 *var2csa*, all assembled transcripts significantly annotated as such were identified. For each
892 conserved gene, Salmon normalised read counts (adjusted for life cycle stage) were summed and
893 expression compared across the generations using a pairwise Wilcoxon rank test.
894
895 Differential expression of *var* domains from *ex vivo* to *in vitro* samples
896 Domain expression was quantified using featureCounts, as described above (Liao *et al.*, 2014).
897 DESeq2 was used to test for differential domain expression, with five expected read counts in at
898 least three patient isolates required, with life cycle stage and patient identity used as covariates.
899 For the *ex vivo* versus *in vitro* comparisons, only *ex vivo* samples that had paired samples in
900 generation 1 underwent differential expression analysis, given the extreme nature of the
901 polymorphism seen in the *var* genes.
902
903 *Var* group expression analysis
904 The type of the *var* gene is determined by multiple parameters: upstream sequence (ups),
905 chromosomal location, direction of transcription and domain composition. All regular *var* genes
906 encode a DBLα domain in the N-terminus of the PfEMP1 protein (Figure 1c). The type of this
907 domain correlates with previously defined *var* gene groups, with group A encoding DBLα1, groups
908 B and C encoding DBLα0 and group B encoding a DBLα2 (chimera between DBLα0 and DBLα1)
909 (Figure 1c). The DBLα domain sequence for each transcript was determined and for each patient
910 a reference of all assembled DBLα domains combined. The relevant sample's non-core reads were
911 mapped to this using Salmon and DBLα expression quantified (Patro *et al.*, 2017). DESeq2
912 normalisation was performed, with patient identity and life cycle stage proportions included as
913 covariates and differences in the amounts of *var* transcripts of group A compared with groups B
914 and C assessed (Love *et al.*, 2014). A similar approach was repeated for NTS domains. NTSA
915 domains are found encoded in group A *var* genes and NTSB domains are found encoded in group
916 B and C *var* genes (Figure 1c).

917
918 Quantification of total *var* gene expression
919 The RNA-sequencing reads were blastn (with the short-blastn option on and significance = 1e-10)
920 against the LARSFADIG nucleotide sequences (142 unique LARSFADIG sequences) to identify reads
921 containing the LARSFADIG motifs. This approach has been described previously (Andrade *et al.*,
922 2020). Once the reads containing the LARSFADIG motifs had been identified, they were used to
923 assembled the LARSFADIG motif. Trinity (Henschel, 2012) and rnaSPAdes (Bushmanova *et al.*,
924 2019) were used separately to assemble the LARSFADIG motif, and the results compared. The
925 sequencing reads were mapped back against the assemblies using bwa mem (Li, 2013), parameter
926 -k 31 -a (as in Andrade *et al.*, 2020). Coverage over the LARSFADIG motif was assessed by
927 determining the coverage over the middle of the motif (S) using Samtools depth (Danecek *et al.*,
928 2021). These values were divided by the number of reads mapped to the *var* exon 1 database and
929 the 3D7 genome (which had *var* genes removed) to represent the proportion of total gene
930 expression dedicated to *var* gene expression (similar to an RPKM). The results of both approaches
931 were compared. This method has been validated on 3D7, IT and HB3 *Plasmodium* strains. *Var2csa*
932 does not contain the LARSFADIG motif, hence this quantitative analysis of global *var* gene
933 expression excluded *var2csa* (which was analysed separately). Significant differences in total *var*
934 gene expression were tested by constructing a linear model with the proportion of gene
935 expression dedicated to *var* gene expression as the response variable, the generation and life cycle
936 stage as an independent variables and the patient identity included as a random effect.
937
938 *Var* expression profiling by DBLα-tag sequencing
939 DBLα-tag sequence analysis was performed as in the original analysis (Wichers *et al.*, 2021), with
940 Varia used to predict domain composition (Mackenzie *et al.*, 2022). The proportion of transcripts
941 encoding NTSA, NTSB, DBLα1, DBLα2 and DBLα0 domains were determined for each sample.
942 These expression levels were used as an alternative approach to see whether there were changes
943 in the *var* group expression levels through culture.
944
945 The consistency of domain annotations was also investigated between the DBLα-tag approach and
946 the assembled transcripts. This was investigated on a per patient basis, with all the predicted
947 annotations from the DBLα-tag approach for a given patient combined. These were compared to
948 the annotations from all assembled transcripts for a given patient. DBLα annotations and DBLα-
949 CIDR annotations were compared. This provided another validation of the whole transcript
950 approach after the domain annotation step and was not dependent on performing differential
951 expression analysis.
952
953 For comparison of both approaches (DBLα-tag sequencing and our new whole transcript
954 approach), the same analysis was performed as in the original analysis (Wichers *et al.*, 2021). All
955 conserved variants (*var1, var2csa* and *var3*) were removed as they were not properly amplified by
956 the DBLα-tag approach. To identify how many assembled transcripts, specifically the DBLα region,
957 were found in the DBLα-tag approach, we applied BLAST. As in the original analysis, a BLAST
958 database was created from the DBLα-tag cluster results and screened for the occurrence of those
959 assembled DBLα regions with more than 97% seq id using the "megablast" option. This was
960 restricted to the assembled DBLα regions that were expressed in the top 75th percentile to allow
961 for a fair comparison, as only DBLα-tag clusters with more than 10 reads were considered.
962 Similarly, to identify how many DBLα-tag sequences were found in the assembled transcripts, a
963 BLAST database was created from the assembled transcripts and screened for the occurrence of
964 the DBLα-tag sequences with more than 97% seq id using the "megablast" option. This was
965 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 #323759012) 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

Almelli, T., Nuel, G., Bischoff, E., Aubouy, A., Elati, M., Wang, C.W., Dillies, M.A., Coppee, J.Y., Ayissi, G.N., Basco, L.K., Rogier, C., Ndam, N.T., Deloron, P., and Tahar, R. (2014) Differences in gene transcriptomic pattern of Plasmodium falciparum in children with cerebral malaria and asymptomatic carriers. *PLoS One* **9**: e114401.

Anders, S., Pyl, P.T., and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.

1014 Andrade, C.M., Fleckenstein, H., Thomson-Luque, R., Doumbo, S., Lima, N.F., Anderson, C.,
1015 Hibbert, J., Hopp, C.S., Tran, T.M., Li, S., Niangaly, M., Cisse, H., Doumtabe, D.,
1016 Skinner, J., Sturdevant, D., Ricklefs, S., Virtaneva, K., Asghar, M., Homann, M.V.,
1017 Turner, L., Martins, J., Allman, E.L., N'Dri, M.E., Winkler, V., Llinas, M., Lavazec, C.,
1018 Martens, C., Farnert, A., Kayentao, K., Ongoiba, A., Lavstsen, T., Osorio, N.S., Otto,
1019 T.D., Recker, M., Traore, B., Crompton, P.D., and Portugal, S. (2020) Increased
1020 circulation time of Plasmodium falciparum underlies persistent asymptomatic infection in
1021 the dry season. *Nat Med* **26**: 1929-1940.
1022 Andreadaki, M., Pace, T., Grasso, F., Siden-Kiamos, I., Mochi, S., Picci, L., Bertuccini, L.,
1023 Ponzi, M., and Curra, C. (2020) Plasmodium berghei Gamete Egress Protein is required
1024 for fertility of both genders. *Microbiologyopen* **9**: e1038.
1025 Avril, M., Tripathi, A.K., Brazier, A.J., Andisi, C., Janes, J.H., Soma, V.L., Sullivan, D.J., Jr.,
1026 Bull, P.C., Stins, M.F., and Smith, J.D. (2012) A restricted subset of var genes mediates
1027 adherence of Plasmodium falciparum-infected erythrocytes to brain endothelial cells.
1028 *Proc Natl Acad Sci U S A* **109**: E1782-1790.
1029 Bachmann, A., Predehl, S., May, J., Harder, S., Burchard, G.D., Gilberger, T.W., Tannich, E.,
1030 and Bruchhaus, I. (2011) Highly co-ordinated var gene expression and switching in
1031 clinical Plasmodium falciparum isolates from non-immune malaria patients. *Cell*
1032 *Microbiol* **13**: 1397-1409.
1033 Baruch, D.I., Pasloske, B.L., Singh, H.B., Bi, X., Ma, X.C., Feldman, M., Taraschi, T.F., and
1034 Howard, R.J. (1995) Cloning the P. falciparum gene encoding PfEMP1, a malarial
1035 variant antigen and adherence receptor on the surface of parasitized human erythrocytes.
1036 *Cell* **82**: 77-87.
1037 Beeson, J.G., Drew, D.R., Boyle, M.J., Feng, G., Fowkes, F.J., and Richards, J.S. (2016)
1038 Merozoite surface proteins in red blood cell invasion, immunity and vaccines against
1039 malaria. *FEMS Microbiol Rev* **40**: 343-372.
1040 Bernabeu, M., Danziger, S.A., Avril, M., Vaz, M., Babar, P.H., Brazier, A.J., Herricks, T., Maki,
1041 J.N., Pereira, L., Mascarenhas, A., Gomes, E., Chery, L., Aitchison, J.D., Rathod, P.K.,
1042 and Smith, J.D. (2016) Severe adult malaria is associated with specific PfEMP1 adhesion
1043 types and high parasite biomass. *Proc Natl Acad Sci U S A* **113**: E3270-3279.
1044 Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011) Scaffolding pre-
1045 assembled contigs using SSPACE. *Bioinformatics* **27**: 578-579.
1046 Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. (2003) The
1047 transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.
1048 *PLoS Biol* **1**: E5.
1049 Brown, A.C., and Guler, J.L. (2020) From Circulation to Cultivation: Plasmodium In Vivo
1050 versus In Vitro. *Trends Parasitol* **36**: 914-926.
1051 Bruske, E.I., Dimonte, S., Enderes, C., Tschan, S., Flotenmeyer, M., Koch, I., Berger, J.,
1052 Kremsner, P., and Frank, M. (2016) In Vitro Variant Surface Antigen Expression in
1053 Plasmodium falciparum Parasites from a Semi-Immune Individual Is Not Correlated with
1054 Var Gene Transcription. *PLoS One* **11**: e0166135.
1055 Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019) rnaSPAdes: a de novo
1056 transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**.
1057 Carrington, E., Cooijmans, R.H.M., Keller, D., Toenhake, C.G., Bartfai, R., and Voss, T.S.
1058 (2021) The ApiAP2 factor PfAP2-HC is an integral component of heterochromatin in the
1059 malaria parasite Plasmodium falciparum. *iScience* **24**: 102444.
1060 Claessens, A., Adams, Y., Ghumra, A., Lindergard, G., Buchan, C.C., Andisi, C., Bull, P.C.,
1061 Mok, S., Gupta, A.P., Wang, C.W., Turner, L., Arman, M., Raza, A., Bozdech, Z., and
1062 Rowe, J.A. (2012) A subset of group A-like var genes encodes the malaria parasite
1063 ligands for binding to human brain endothelial cells. *Proc Natl Acad Sci U S A* **109**:
1064 E1772-1781.

1065 Claessens, A., Affara, M., Assefa, S.A., Kwiatkowski, D.P., and Conway, D.J. (2017) Culture
1066 adaptation of malaria parasites selects for convergent loss-of-function mutants. *Sci Rep* **7**:
1067 41303.
1068 Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A.,
1069 Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021) Twelve years of SAMtools
1070 and BCFtools. *Gigascience* **10**.
1071 Dimonte, S., Bruske, E.I., Hass, J., Supan, C., Salazar, C.L., Held, J., Tschan, S., Esen, M.,
1072 Flotenmeyer, M., Koch, I., Berger, J., Bachmann, A., Sim, B.K., Hoffman, S.L.,
1073 Kremsner, P.G., Mordmuller, B., and Frank, M. (2016) Sporozoite Route of Infection
1074 Influences In Vitro var Gene Transcription of Plasmodium falciparum Parasites From
1075 Controlled Human Infections. *J Infect Dis* **214**: 884-894.
1076 Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F., and Jiang, Y. (2021) SCDC:
1077 bulk gene expression deconvolution by multiple single-cell RNA sequencing references.
1078 *Brief Bioinform* **22**: 416-427.
1079 Finn, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: interactive sequence
1080 similarity searching. *Nucleic Acids Res* **39**: W29-37.
1081 Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-
1082 generation sequencing data. *Bioinformatics* **28**: 3150-3152.
1083 Ge, S.X., Jung, D., and Yao, R. (2020) ShinyGO: a graphical gene-set enrichment tool for
1084 animals and plants. *Bioinformatics* **36**: 2628-2629.
1085 Guillochon, E., Fraering, J., Joste, V., Kamaliddin, C., Vianou, B., Houze, L., Baudrin, L.G.,
1086 Faucher, J.F., Aubouy, A., Houze, S., Cot, M., Argy, N., Taboureau, O., Bertin, G.I., and
1087 Neuro, C.M.g. (2022) Transcriptome Analysis of Plasmodium falciparum Isolates From
1088 Benin Reveals Specific Gene Expression Associated With Cerebral Malaria. *J Infect Dis*
1089 **225**: 2187-2196.
1090 Henschel, R., Lieber, M., Wu, L., Nista, P. M., Haas, B. J., LeDuc, R. D., (2012) Trinity RNA-
1091 Seq assembler performance optimization. In: XSEDE '12: Proceedings of the 1st
1092 Conference of the Extreme Science and Engineering Discovery Environment: Bridging
1093 from the eXtreme to the campus and beyond. ACM, pp. 1-8.
1094 Hoo, R., Bruske, E., Dimonte, S., Zhu, L., Mordmuller, B., Sim, B.K.L., Kremsner, P.G.,
1095 Hoffman, S.L., Bozdech, Z., Frank, M., and Preiser, P.R. (2019) Transcriptome profiling
1096 reveals functional variation in Plasmodium falciparum parasites from controlled human
1097 malaria infection studies. *EBioMedicine* **48**: 442-452.
1098 Howick, V.M., Russell, A.J.C., Andrews, T., Heaton, H., Reid, A.J., Natarajan, K., Butungi, H.,
1099 Metcalf, T., Verzier, L.H., Rayner, J.C., Berriman, M., Herren, J.K., Billker, O.,
1100 Hemberg, M., Talman, A.M., and Lawniczak, M.K.N. (2019) The Malaria Cell Atlas:
1101 Single parasite transcriptomes across the complete Plasmodium life cycle. *Science* **365**.
1102 Huang, X., and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res* **9**:
1103 868-877.
1104 Jensen, A.T., Magistrado, P., Sharp, S., Joergensen, L., Lavstsen, T., Chiucchiuini, A., Salanti,
1105 A., Vestergaard, L.S., Lusingu, J.P., Hermsen, R., Sauerwein, R., Christensen, J.,
1106 Nielsen, M.A., Hviid, L., Sutherland, C., Staalsoe, T., and Theander, T.G. (2004)
1107 Plasmodium falciparum associated with severe childhood malaria preferentially
1108 expresses PfEMP1 encoded by group A var genes. *J Exp Med* **199**: 1179-1190.
1109 Jespersen, J.S., Wang, C.W., Mkumbaye, S.I., Minja, D.T., Petersen, B., Turner, L., Petersen,
1110 J.E., Lusingu, J.P., Theander, T.G., and Lavstsen, T. (2016) Plasmodium falciparum var
1111 genes expressed in children with severe malaria encode CIDRalpha1 domains. *EMBO*
1112 *Mol Med* **8**: 839-850.
1113 Joste, V., Guillochon, E., Fraering, J., Vianou, B., Watier, L., Jafari-Guemouri, S., Cot, M.,
1114 Houze, S., Aubouy, A., Faucher, J.F., Argy, N., and Bertin, G.I. (2020) PfEMP1 A-Type
1115 ICAM-1-Binding Domains Are Not Associated with Cerebral Malaria in Beninese
1116 Children. *mBio* **11**.

1117 Kaneko, I., Iwanaga, S., Kato, T., Kobayashi, I., and Yuda, M. (2015) Genome-Wide
1118 Identification of the Target Genes of AP2-O, a Plasmodium AP2-Family Transcription
1119 Factor. *PLoS Pathog* **11**: e1004905.
1120 Kessler, A., Dankwa, S., Bernabeu, M., Harawa, V., Danziger, S.A., Duffy, F., Kampondeni,
1121 S.D., Potchen, M.J., Dambrauskas, N., Vigdorovich, V., Oliver, B.G., Hochman, S.E.,
1122 Mowrey, W.B., MacCormick, I.J.C., Mandala, W.L., Rogerson, S.J., Sather, D.N.,
1123 Aitchison, J.D., Taylor, T.E., Seydel, K.B., Smith, J.D., and Kim, K. (2017) Linking
1124 EPCR-Binding PfEMP1 to Brain Swelling in Pediatric Cerebral Malaria. *Cell Host*
1125 *Microbe* **22**: 601-614 e605.
1126 Kirchgatter, K., and Portillo Hdel, A. (2002) Association of severe noncerebral Plasmodium
1127 falciparum malaria in Brazil with expressed PfEMP1 DBL1 alpha sequences lacking
1128 cysteine residues. *Mol Med* **8**: 16-23.
1129 Kraemer, S.M., and Smith, J.D. (2003) Evidence for the importance of genetic structuring to the
1130 structural and functional specialization of the Plasmodium falciparum var gene family.
1131 *Mol Microbiol* **50**: 1527-1538.
1132 Kyes, S.A., Kraemer, S.M., and Smith, J.D. (2007) Antigenic variation in Plasmodium
1133 falciparum: gene organization and regulation of the var multigene family. *Eukaryot Cell*
1134 **6**: 1511-1520.
1135 Lavstsen, T., Magistrado, P., Hermsen, C.C., Salanti, A., Jensen, A.T., Sauerwein, R., Hviid, L.,
1136 Theander, T.G., and Staalsoe, T. (2005) Expression of Plasmodium falciparum
1137 erythrocyte membrane protein 1 in experimentally infected humans. *Malar J* **4**: 21.
1138 Lavstsen, T., Salanti, A., Jensen, A.T., Arnot, D.E., and Theander, T.G. (2003) Sub-grouping of
1139 Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-
1140 coding regions. *Malar J* **2**: 27.
1141 Lavstsen, T., Turner, L., Saguti, F., Magistrado, P., Rask, T.S., Jespersen, J.S., Wang, C.W.,
1142 Berger, S.S., Baraka, V., Marquard, A.M., Seguin-Orlando, A., Willerslev, E., Gilbert,
1143 M.T., Lusingu, J., and Theander, T.G. (2012) Plasmodium falciparum erythrocyte
1144 membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in
1145 children. *Proc Natl Acad Sci U S A* **109**: E1791-1800.
1146 Lee, H.J., Georgiadou, A., Walther, M., Nwakanma, D., Stewart, L.B., Levin, M., Otto, T.D.,
1147 Conway, D.J., Coin, L.J., and Cunnington, A.J. (2018) Integrated pathogen load and dual
1148 transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Sci*
1149 *Transl Med* **10**.
1150 Leech, J.H., Barnwell, J.W., Miller, L.H., and Howard, R.J. (1984) Identification of a strain-
1151 specific malarial antigen exposed on the surface of Plasmodium falciparum-infected
1152 erythrocytes. *J Exp Med* **159**: 1567-1575.
1153 Lewis, T.E., Sillitoe, I., and Lees, J.G. (2019) cath-resolve-hits: a new tool that resolves domain
1154 matches suspiciously quickly. *Bioinformatics* **35**: 1766-1767.
1155 Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1156 *arXiv*: 1303.3997v1301 [q-bio.GN].
1157 Liao, Y., Smyth, G.K., and Shi, W. (2013) The Subread aligner: fast, accurate and scalable read
1158 mapping by seed-and-vote. *Nucleic Acids Res* **41**: e108.
1159 Liao, Y., Smyth, G.K., and Shi, W. (2014) featureCounts: an efficient general purpose program
1160 for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.
1161 Lischer, H.E.L., and Shimizu, K.K. (2017) Reference-guided de novo assembly approach
1162 improves genome reconstruction for related species. *BMC Bioinformatics* **18**: 474.
1163 Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and
1164 dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.
1165 Mackenzie, G., Jensen, R.W., Lavstsen, T., and Otto, T.D. (2022) Varia: a tool for prediction,
1166 analysis and visualisation of variable genes. *BMC Bioinformatics* **23**: 52.

1167 Mackinnon, M.J., Li, J., Mok, S., Kortok, M.M., Marsh, K., Preiser, P.R., and Bozdech, Z.
1168 (2009) Comparative transcriptional and genomic analysis of Plasmodium falciparum
1169 field isolates. *PLoS Pathog* **5**: e1000644.
1170 MalariaGen, Ahouidi, A., Ali, M., Almagro-Garcia, J., Amambua-Ngwa, A., Amaratunga, C.,
1171 Amato, R., Amenga-Etego, L., Andagalu, B., Anderson, T.J.C., Andrianaranjaka, V.,
1172 Apinjoh, T., Ariani, C., Ashley, E.A., Auburn, S., Awandare, G.A., Ba, H., Baraka, V.,
1173 Barry, A.E., Bejon, P., Bertin, G.I., Boni, M.F., Borrmann, S., Bousema, T., Branch, O.,
1174 Bull, P.C., Busby, G.B.J., Chookajorn, T., Chotivanich, K., Claessens, A., Conway, D.,
1175 Craig, A., D'Alessandro, U., Dama, S., Day, N.P., Denis, B., Diakite, M., Djimde, A.,
1176 Dolecek, C., Dondorp, A.M., Drakeley, C., Drury, E., Duffy, P., Echeverry, D.F.,
1177 Egwang, T.G., Erko, B., Fairhurst, R.M., Faiz, A., Fanello, C.A., Fukuda, M.M.,
1178 Gamboa, D., Ghansah, A., Golassa, L., Goncalves, S., Hamilton, W.L., Harrison, G.L.A.,
1179 Hart, L., Henrichs, C., Hien, T.T., Hill, C.A., Hodgson, A., Hubbart, C., Imwong, M.,
1180 Ishengoma, D.S., Jackson, S.A., Jacob, C.G., Jeffery, B., Jeffreys, A.E., Johnson, K.J.,
1181 Jyothi, D., Kamaliddin, C., Kamau, E., Kekre, M., Kluczynski, K., Kochakarn, T.,
1182 Konate, A., Kwiatkowski, D.P., Kyaw, M.P., Lim, P., Lon, C., Loua, K.M., Maiga-
1183 Ascofare, O., Malangone, C., Manske, M., Marfurt, J., Marsh, K., Mayxay, M., Miles,
1184 A., Miotto, O., Mobegi, V., Mokuolu, O.A., Montgomery, J., Mueller, I., Newton, P.N.,
1185 Nguyen, T., Nguyen, T.N., Noedl, H., Nosten, F., Noviyanti, R., Nzila, A.*, et al.* (2021)
1186 An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide
1187 samples. *Wellcome Open Res* **6**: 42.
1188 Mkumbaye, S.I., Wang, C.W., Lyimo, E., Jespersen, J.S., Manjurano, A., Mosha, J., Kavishe,
1189 R.A., Mwakalinga, S.B., Minja, D.T.R., Lusingu, J.P., Theander, T.G., and Lavstsen, T.
1190 (2017) The Severity of Plasmodium falciparum Infection Is Associated with Transcript
1191 Levels of var Genes Encoding Endothelial Protein C Receptor-Binding P. falciparum
1192 Erythrocyte Membrane Protein 1. *Infect Immun* **85**.
1193 Mok, B.W., Ribacke, U., Rasti, N., Kironde, F., Chen, Q., Nilsson, P., and Wahlgren, M. (2008)
1194 Default Pathway of var2csa switching and translational repression in Plasmodium
1195 falciparum. *PLoS One* **3**: e1982.
1196 Oakley, M.S., Kumar, S., Anantharaman, V., Zheng, H., Mahajan, B., Haynes, J.D., Moch, J.K.,
1197 Fairhurst, R., McCutchan, T.F., and Aravind, L. (2007) Molecular factors and
1198 biochemical pathways induced by febrile temperature in intraerythrocytic Plasmodium
1199 falciparum parasites. *Infect Immun* **75**: 2012-2025.
1200 Otto, T.D., Assefa, S.A., Bohme, U., Sanders, M.J., Kwiatkowski, D., Pf3k, c., Berriman, M.,
1201 and Newbold, C. (2019) Evolutionary analysis of the most polymorphic gene family in
1202 falciparum malaria. *Wellcome Open Res* **4**: 193.
1203 Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017) Salmon provides fast
1204 and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417-419.
1205 Peters, J.M., Fowler, E.V., Krause, D.R., Cheng, Q., and Gatton, M.L. (2007) Differential
1206 changes in Plasmodium falciparum var transcription during adaptation to culture. *J Infect*
1207 *Dis* **195**: 748-755.
1208 Pickford, A.K., Michel-Todo, L., Dupuy, F., Mayor, A., Alonso, P.L., Lavazec, C., and Cortes,
1209 A. (2021) Expression Patterns of Plasmodium falciparum Clonally Variant Genes at the
1210 Onset of a Blood Infection in Malaria-Naive Humans. *mBio* **12**: e0163621.
1211 Quintana, M.D.P., Ecklu-Mensah, G., Tcherniuk, S.O., Ditlev, S.B., Oleinikov, A.V., Hviid, L.,
1212 and Lopez-Perez, M. (2019) Comprehensive analysis of Fc-mediated IgM binding to the
1213 Plasmodium falciparum erythrocyte membrane protein 1 family in three parasite clones.
1214 *Sci Rep* **9**: 6050.
1215 Rask, T.S., Hansen, D.A., Theander, T.G., Gorm Pedersen, A., and Lavstsen, T. (2010)
1216 Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--
1217 divide and conquer. *PLoS Comput Biol* **6**.

1218 Robert, F., Ntoumi, F., Angel, G., Candito, D., Rogier, C., Fandeur, T., Sarthou, J.L., and
1219 Mercereau-Puijalon, O. (1996) Extensive genetic diversity of Plasmodium falciparum
1220 isolates collected from patients with severe malaria in Dakar, Senegal. *Trans R Soc Trop*
1221 *Med Hyg* **90**: 704-711.
1222 Rorick, M.M., Rask, T.S., Baskerville, E.B., Day, K.P., and Pascual, M. (2013) Homology
1223 blocks of Plasmodium falciparum var genes and clinically distinct forms of severe
1224 malaria in a local population. *BMC Microbiol* **13**: 244.
1225 Sahu, P.K., Duffy, F.J., Dankwa, S., Vishnyakova, M., Majhi, M., Pirpamer, L., Vigdorovich,
1226 V., Bage, J., Maharana, S., Mandala, W., Rogerson, S.J., Seydel, K.B., Taylor, T.E.,
1227 Kim, K., Sather, D.N., Mohanty, A., Mohanty, R.R., Mohanty, A., Pattnaik, R.,
1228 Aitchison, J.D., Hoffmann, A., Mohanty, S., Smith, J.D., Bernabeu, M., and Wassmer,
1229 S.C. (2021) Determinants of brain swelling in pediatric and adult cerebral malaria. *JCI*
1230 *Insight* **6**.
1231 Salanti, A., Dahlback, M., Turner, L., Nielsen, M.A., Barfod, L., Magistrado, P., Jensen, A.T.,
1232 Lavstsen, T., Ofori, M.F., Marsh, K., Hviid, L., and Theander, T.G. (2004) Evidence for
1233 the involvement of VAR2CSA in pregnancy-associated malaria. *J Exp Med* **200**: 1197-
1234 1203.
1235 Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J., and
1236 Lanzer, M. (1998) Antigenic variation in malaria: in situ switching, relaxed and mutually
1237 exclusive transcription of var genes during intra-erythrocytic development in
1238 Plasmodium falciparum. *EMBO J* **17**: 5418-5426.
1239 Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012) Oases: robust de novo RNA-
1240 seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**: 1086-
1241 1092.
1242 Shabani, E., Hanisch, B., Opoka, R.O., Lavstsen, T., and John, C.C. (2017) Plasmodium
1243 falciparum EPCR-binding PfEMP1 expression increases with malaria disease severity
1244 and is elevated in retinopathy negative cerebral malaria. *BMC Med* **15**: 183.
1245 Smith, J.D., Chitnis, C.E., Craig, A.G., Roberts, D.J., Hudson-Taylor, D.E., Peterson, D.S.,
1246 Pinches, R., Newbold, C.I., and Miller, L.H. (1995) Switches in expression of
1247 Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent
1248 phenotypes of infected erythrocytes. *Cell* **82**: 101-110.
1249 Stevenson, L., Laursen, E., Cowan, G.J., Bandoh, B., Barfod, L., Cavanagh, D.R., Andersen,
1250 G.R., and Hviid, L. (2015) alpha2-Macroglobulin Can Crosslink Multiple Plasmodium
1251 falciparum Erythrocyte Membrane Protein 1 (PfEMP1) Molecules and May Facilitate
1252 Adhesion of Parasitized Erythrocytes. *PLoS Pathog* **11**: e1005022.
1253 Storm, J., Jespersen, J.S., Seydel, K.B., Szestak, T., Mbewe, M., Chisala, N.V., Phula, P., Wang,
1254 C.W., Taylor, T.E., Moxon, C.A., Lavstsen, T., and Craig, A.G. (2019) Cerebral malaria
1255 is associated with differential cytoadherence to brain endothelial cells. *EMBO Mol Med*
1256 **11**.
1257 Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch,
1258 J.A., and Wellems, T.E. (1995) The large diverse gene family var encodes proteins
1259 involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected
1260 erythrocytes. *Cell* **82**: 89-100.
1261 Tarr, S.J., Diaz-Ingelmo, O., Stewart, L.B., Hocking, S.E., Murray, L., Duffy, C.W., Otto, T.D.,
1262 Chappell, L., Rayner, J.C., Awandare, G.A., and Conway, D.J. (2018) Schizont
1263 transcriptome variation among clinical isolates and laboratory-adapted clones of the
1264 malaria parasite Plasmodium falciparum. *BMC Genomics* **19**: 894.
1265 Taylor, H.M., Grainger, M., and Holder, A.A. (2002) Variation in the expression of a
1266 Plasmodium falciparum protein family implicated in erythrocyte invasion. *Infect Immun*
1267 **70**: 5779-5789.
1268 Tebben, K., Dia, A., and Serre, D. (2022) Determination of the Stage Composition of
1269 Plasmodium Infections from Bulk Gene Expression Data. *mSystems* **7**: e0025822.

1270 Tonkin-Hill, G.Q., Trianty, L., Noviyanti, R., Nguyen, H.H.T., Sebayang, B.F., Lampah, D.A.,
1271 Marfurt, J., Cobbold, S.A., Rambhatla, J.S., McConville, M.J., Rogerson, S.J., Brown,
1272 G.V., Day, K.P., Price, R.N., Anstey, N.M., Papenfuss, A.T., and Duffy, M.F. (2018) The
1273 Plasmodium falciparum transcriptome in severe malaria reveals altered expression of
1274 genes involved in important processes including surface antigen-encoding var genes.
1275 *PLoS Biol* **16**: e2004328.
1276 Tuikue Ndam, N., Moussiliou, A., Lavstsen, T., Kamaliddin, C., Jensen, A.T.R., Mama, A.,
1277 Tahar, R., Wang, C.W., Jespersen, J.S., Alao, J.M., Gamain, B., Theander, T.G., and
1278 Deloron, P. (2017) Parasites Causing Cerebral Falciparum Malaria Bind Multiple
1279 Endothelial Receptors and Express EPCR and ICAM-1-Binding PfEMP1. *J Infect Dis*
1280 **215**: 1918-1925.
1281 Turner, L., Lavstsen, T., Berger, S.S., Wang, C.W., Petersen, J.E., Avril, M., Brazier, A.J.,
1282 Freeth, J., Jespersen, J.S., Nielsen, M.A., Magistrado, P., Lusingu, J., Smith, J.D.,
1283 Higgins, M.K., and Theander, T.G. (2013) Severe malaria is associated with parasite
1284 binding to endothelial protein C receptor. *Nature* **498**: 502-505.
1285 Ukaegbu, U.E., Kishore, S.P., Kwiatkowski, D.L., Pandarinath, C., Dahan-Pasternak, N.,
1286 Dzikowski, R., and Deitsch, K.W. (2014) Recruitment of PfSET2 by RNA polymerase II
1287 to variant antigen encoding loci contributes to antigenic variation in P. falciparum. *PLoS*
1288 *Pathog* **10**: e1003854.
1289 Ukaegbu, U.E., Zhang, X., Heinberg, A.R., Wele, M., Chen, Q., and Deitsch, K.W. (2015) A
1290 Unique Virulence Gene Occupies a Principal Position in Immune Evasion by the Malaria
1291 Parasite Plasmodium falciparum. *PLoS Genet* **11**: e1005234.
1292 Vignali, M., Armour, C.D., Chen, J., Morrison, R., Castle, J.C., Biery, M.C., Bouzek, H., Moon,
1293 W., Babak, T., Fried, M., Raymond, C.K., and Duffy, P.E. (2011) NSR-seq
1294 transcriptional profiling enables identification of a gene signature of Plasmodium
1295 falciparum parasites infecting children. *J Clin Invest* **121**: 1119-1129.
1296 Wagner, B.D., Grunwald, G.K., Zerbe, G.O., Mikulich-Gilbertson, S.K., Robertson, C.E.,
1297 Zemanick, E.T., and Harris, J.K. (2018) On the Use of Diversity Measures in
1298 Longitudinal Sequencing Studies of Microbial Communities. *Front Microbiol* **9**: 1037.
1299 Wahlgren, M., Goel, S., and Akhouri, R.R. (2017) Variant surface antigens of Plasmodium
1300 falciparum and their roles in severe malaria. *Nat Rev Microbiol* **15**: 479-491.
1301 Wang, W., Chen, E.Z., and Li, H. (2021) Truncated Rank-Based Tests for Two-Part Models with
1302 Excessive Zeros and Applications to Microbiome Data. *arXiv*.
1303 Warimwe, G.M., Recker, M., Kiragu, E.W., Buckee, C.O., Wambua, J., Musyoki, J.N., Marsh,
1304 K., and Bull, P.C. (2013) Plasmodium falciparum var gene expression homogeneity as a
1305 marker of the host-parasite relationship under different levels of naturally acquired
1306 immunity to malaria. *PLoS One* **8**: e70467.
1307 WHO, (2022) World malaria report 2022. In. Geneva, pp.
1308 Wichers, J.S., Scholz, J.A.M., Strauss, J., Witt, S., Lill, A., Ehnold, L.I., Neupert, N., Liffner, B.,
1309 Luhken, R., Petter, M., Lorenzen, S., Wilson, D.W., Low, C., Lavazec, C., Bruchhaus, I.,
1310 Tannich, E., Gilberger, T.W., and Bachmann, A. (2019) Dissecting the Gene Expression,
1311 Localization, Membrane Topology, and Function of the Plasmodium falciparum
1312 STEVOR Protein Family. *mBio* **10**.
1313 Wichers, J.S., Tonkin-Hill, G., Thye, T., Krumkamp, R., Kreuels, B., Strauss, J., von Thien, H.,
1314 Scholz, J.A., Smedegaard Hansson, H., Weisel Jensen, R., Turner, L., Lorenz, F.R.,
1315 Schollhorn, A., Bruchhaus, I., Tannich, E., Fendel, R., Otto, T.D., Lavstsen, T.,
1316 Gilberger, T.W., Duffy, M.F., and Bachmann, A. (2021) Common virulence gene
1317 expression in adult first-time infected malaria patients and severe cases. *Elife* **10**.
1318 Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou,
1319 X., Lam, T.W., Li, Y., Xu, X., Wong, G.K., and Wang, J. (2014) SOAPdenovo-Trans: de
1320 novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**: 1660-1666.

1321 Yamagishi, J., Natori, A., Tolba, M.E., Mongan, A.E., Sugimoto, C., Katayama, T., Kawashima,
1322 S., Makalowski, W., Maeda, R., Eshita, Y., Tuda, J., and Suzuki, Y. (2014) Interactive
1323 transcriptome analysis of malaria patients and infecting Plasmodium falciparum. *Genome*
1324 *Res* **24**: 1433-1444.
1325 Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M.,
1326 Zody, M.C., and Henn, M.R. (2012) De novo assembly of highly diverse viral
1327 populations. *BMC Genomics* **13**: 475.
1328 Yuda, M., Iwanaga, S., Shigenobu, S., Kato, T., and Kaneko, I. (2010) Transcription factor AP2-
1329 Sp and its target genes in malarial sporozoites. *Mol Microbiol* **75**: 854-863.
1330 Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using
1331 de Bruijn graphs. *Genome Res* **18**: 821-829.
1332 Zhang, Q., Zhang, Y., Huang, Y., Xue, X., Yan, H., Sun, X., Wang, J., McCutchan, T.F., and
1333 Pan, W. (2011) From in vivo to in vitro: dynamic analysis of Plasmodium falciparum var
1334 gene expression patterns of patient isolates during adaptation to culture. *PLoS One* **6**:
1335 e20591.
1336 Zhang, X., Florini, F., Visone, J.E., Lionardi, I., Gross, M.R., Patel, V., and Deitsch, K.W.
1337 (2022) A coordinated transcriptional switching network mediates antigenic variation of
1338 human malaria parasites. *Elife* **11**.
1339 Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013) The
1340 MaSuRCA genome assembler. *Bioinformatics* **29**: 2669-2677.
1341

1342 Figure Legends

1343

1344 **Figure 1: Summary of the *var* chromosomal location, *var* gene, PfEMP1 protein structure, and**
1345 **PfEMP1 binding phenotypes. a)** Chromosomal position and transcriptional direction (indicated
1346 by arrows) of the different *var* gene groups, designated by the respective type of upstream
1347 sequence (Kraemer and Smith et al., 2003, Lavstsen et al., 2003). **b)** Structure of the *var* gene
1348 which encodes the PfEMP1 protein. The *var* gene is composed of two exons, the first, around 3–
1349 9.4 kb, encodes the highly variable extracellular region and the transmembrane region (TM) of
1350 PfEMP1. Exon 2 is shorter with about 1.2 kb and encodes a semi-conserved intracellular region
1351 (acidic terminal segment, ATS). The PfEMP1 protein is composed of an N-terminal segment (NTS),
1352 followed by a variable number of Duffy binding-like (DBL) domains and cysteine-rich interdomain
1353 regions (CIDR) (Rask et al., 2010). **c)** Summary of PfEMP1 proteins encoded in the parasite
1354 genome, their composition of domain subtypes and associated N-terminal binding phenotype.
1355 Group A and some B proteins have an EPCR-binding phenotype; the vast majority of group B and
1356 C PfEMP1 proteins bind to CD36. Group A proteins also include those that bind a yet unknown
1357 receptor, as well as VAR1 and VAR3 variants with unknown function and binding phenotype.
1358 VAR2CSA (group E) binds placental CSA.
1359

1360 **Figure 2: Overview of novel computational pipelines for assembling *var* transcripts.** The
1361 original approach (green) used SoapDeNovo-Trans (k=71) to perform whole *var* transcript
1362 assembly. The whole transcript approach (blue) focused on assembling whole *var* transcripts from
1363 the non-core reads using rnaSPAdes (k = 71). Contigs were then joined into longer transcripts using
1364 SSPACE. The domain approach (orange) assembled *var* domains first and then joined the domains
1365 into whole transcripts. Domains were assembled separately using three different *de novo*
1366 assemblers (SoapDeNovo2, Oases and rnaSPAdes). Next, a reference of assembled domains was
1367 composed and cd-hit (at sequence identity = 99%) was used to remove redundant sequences.
1368 Cap3 was used to merge and extend domain assemblies. Finally, SSPACE was used to join domains
1369 together. HMM models built on the Rask *et al.*, 2010 dataset were used to annotate the assembled
1370 transcripts (Rask *et al.*, 2010). The most significant alignment was taken as the best annotation for

each region of the assembled transcript (significance <= 1e-5) identified using cath-resolve-hits0. Transcripts < 500nt 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).

Figure 2 – Figure supplement 1: Performance of novel computational pipelines for *var* assembly on *Plasmodium falciparum* 3D7: The three approaches (whole transcript: blue, domain approach: orange, original approach: green) were applied to a public RNA-seq dataset (ENA: PRJEB31535) of the intra-erythrocytic life cycle stages of 3 biological replicates of cultured *P. falciparum* 3D7, sampled at 8-hour intervals up until 40hrs post infection (hpi) and then at 4-hour intervals up until 48 hpi (Wichers et al., 2019). Boxplots show the data from the 3 biological replicates for each time point in the intra-erythrocytic life cycle: **a)** alignment scores for the dominantly expressed *var* gene (PF3D7_0712600), **b)** accuracy scores for the dominantly expressed *var* gene (PF3D7_0712600), **c)** number of contigs required to assemble the dominant *var* gene (PF3D7_0712600), **d)** alignment scores for a middle ranking expressed *var* gene (PF3D7_0937800), **e)** alignment scores for the lowest expressed *var* gene (PF3D7_0200100). The first best blast 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 √accuracy* 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. Assembly completeness of the dominant *var* gene (PF3D7_0712600, length = 6648nt) for the three approaches was assessed for each biological replicate: **f)** biological replicate 1, **g)** biological replicate 2, **h)** biological replicate 3. Dotted lines represent the start and end of the contigs required to assemble the *var* gene. Red bars represent assembled sequences relative to the dominantly expressed whole *var* gene sequence, where we know the true sequence (termed "reference transcript").

Figure 2 – Figure supplement 2: Histograms of the length and frequency of *var* transcripts produced by the different *var* assembly approaches. Approaches were applied to *ex vivo* samples (n =32) from Wichers *et al*., 2021. Assembled *var* transcripts < 500nt and not containing a significantly annotated *var* domain were removed.

Figure 2 – Figure supplement 3: Domain expression correlation between the whole transcript approach and the original approach. Correlation plots of the original approach *var* domain counts vs those produced by the whole transcript approach. Each plot represents a patient. Domain counts were determined using featureCounts in both approaches and counts for each domain aggregated. Plots show $\log_2$ transformed normalized counts. The line represents a linear regression model fit.

Figure 2 – Figure supplement 4: *Var* transcript expression differences derived from the whole transcript approach between a) pre-exposed (n=17) (blue) vs naïve (n=15) (orange) cases and b) severe (n=8) (orange) vs non-severe (n=24) (blue) cases. Boxplots show $\log_2$ normalized Salmon counts adjusted for life cycle stage (derived from the mixture model approach). Comparison of *var* transcript expression differences found in the original approach with those assembled using the whole transcript approach. These transcripts were found to be significantly differentially expressed in the **c)** pre-exposed (n=17) (blue) vs naïve (n=15) (orange) cases and **d)** severe (n=8) (orange) vs non-severe (n=24) (blue) cases in the original analysis. Each box represents a comparison of the transcript assembled using the whole transcript approach (grey

1420 background) vs using the original approach (white background). Boxplots show log_2 Salmon
1421 normalized counts. RNA-sequencing reads of each patient sample were matched to *de novo*
1422 assembled *var* contigs with varying lengths and domains using the whole transcript approach.
1423 Shown are significantly differently expressed *var* transcripts. Differential expression analysis was
1424 performed using DESeq2.
1425
1426 **Figure 3: Summary of analyses of *var* and core gene transcriptome changes in paired *ex vivo***
1427 **and short-term *in vitro* cultured parasites.** From a total of 13 parasite isolates, the *ex vivo* samples
1428 (Wichers *et al*., 2021) and the corresponding *in vitro*-cultured parasites of the first (n=13), second
1429 (n=10) and third (n=1) replication cycle were analysed by RNA sequencing. The expression of non-
1430 polymorphic core genes and polymorphic *var* genes was determined in different analysis streams:
1431 (1) Non-polymorphic core gene reads were mapped to the 3D7 reference genome, expression was
1432 quantified using HTSeq and differential expression analysis performed (orange); (2) Non-core
1433 reads were identified, whole transcripts were assembled with rnaSPAdes, expression of both *var*
1434 transcripts (red) and domains (light green) was quantified, and *var* domain differential expression
1435 analysis was performed. "Per patient analysis" (yellow) represents combining all assembled *var*
1436 transcripts for samples originating from the same *ex vivo* sample only. For each conserved *var* gene
1437 (*var1-3D7, var1-IT, var2csa* and *var3*) all significantly assembled conserved *var* transcripts were
1438 identified and put into a combined reference (pink). The normalised counts for each conserved
1439 gene were summed. Non-core reads were mapped to this and DESeq2 normalisation performed.
1440 *Var* type (group A vs group B and C) expression (purple) was quantified using the DBLα and NTS
1441 assembled sequences and differences across generations were assessed. Total *var* gene expression
1442 (turquoise) was quantified by assembling and quantifying the coverage over the highly conserved
1443 LARSFADIG motif, with the performance of assembly using Trinity and rnaSPAdes assessed. DBLα-
1444 tag data was used to confirm the results of the dominant *var* gene expression analysis and the *var*
1445 type analysis (dark green). *Var* expression homogeneity (VEH) was analysed at the patient level (α
1446 diversity curves). All differential expression analyses were performed using DESeq2.To ensure a
1447 fair comparison of samples, which may contain different proportions of life cycle stages, the
1448 performance of two different *in silico* approaches was evaluated by counting Giemsa-stained thin
1449 blood smears (blue).
1450
1451 **Figure 3 – Figure supplement 1: Parasite life cycle stage estimates across the generations.** Life
1452 cycle stage percentages were estimated using two approaches. **a)** The first, termed the mixture
1453 model approach, was the same as used in the original analysis with 8, 19, 30, and 42 hours post
1454 infection corresponding to ring (blue), early trophozoite (orange), late trophozoite (green),
1455 schizont (red) and gametocyte (purple) stages, respectively. **b)** The second approach, termed
1456 SCDC, used SCDC, a deconvolution tool, that uses single cell reference transcriptomes of known
1457 life cycle stages, ring (blue), trophozoite (orange), schizont (red), and gametocyte (purple). The
1458 Malaria Cell Atlas for *P. berghei* was used as the reference transcriptome due to it containing
1459 gametocyte single cells. Only *P. berghei* genes that had 1:1 orthologs with *P. falciparum* 3D7 were
1460 used. Samples are ordered by generation: *ex vivo*; yellow, generation 1; pink, generation 2; blue,
1461 generation 3; green. In panels a and b life cycle stage proportions are shown for all *ex vivo* samples
1462 (n =32), paired generation 1 samples (n=13), paired generation 2 samples (n=10) and a paired
1463 generation 3 Figure shows life cycle stage proportions for all *ex vivo* samples (n =32), paired
1464 generation 1 samples (n=13), paired generation 2 samples (n=10) and a paired generation 3
1465 sample (n=1) **c)** Significant differences in the parasite life cycle stage estimates (derived from the
1466 mixture model approach) for the paired *ex vivo* (n=13), generation 1 (n=13), generation 2 (n=10)

1467 and generation 3 (n=1) samples were tested using paired Wilcox tests (for the ring stage
1468 comparison) and a modified Wilcox on rank test for zero inflated data on the other life cycle stages.
1469 The numbers above the boxplots represent significant adjusted p-values observed (Benjamini-
1470 Hochberg adjusted). **d)** Concordance of life cycle stage estimations using the mixture model
1471 approach and counting Giemsa-stained thin blood smears. Giemsa-stained thin blood smears
1472 counting was performed by two independent researchers. Orange boxplots represent
1473 percentages determined using the mixture model approach and blue boxplots represent
1474 percentages determined using Giemsa-stained thin blood smears. Data shown represents the *ex*
1475 *vivo* (n=11) samples and the paired *in vitro* generations (generation 1; n=12, generation 2; n=9 and
1476 generation 3;n=1) Numbers above the boxplots represent adjusted p-values (Benjamini-Hochberg
1477 adjusted) from a pairwise Wilcoxon test. The same classification as the mixture model approach
1478 was used for the Giemsa-stained thin blood smears counting (8, 19, 30, and 42 hours post infection
1479 corresponding to ring, early trophozoite, late trophozoite and schizont stages respectively).
1480
1481 **Figure 3 – Figure supplement 2: *Var* expression profiles across different mapping approaches.**
1482 Different mapping approaches were used to quantify the *var* expression profiles of each sample
1483 (*ex vivo* (n=13), generation 1 (n=13), generation 2 (n=10) and generation 3 (n=1). The pooled
1484 sample approach in which all significantly assembled *var* transcripts (1500nt and containing 3
1485 significantly annotated *var* domains) across samples were combined into a reference and
1486 redundancy was removed using cd-hit (at sequence identity = 99%) (**a–c**). The non-core reads of
1487 each sample were mapped to this pooled reference using **a)** Salmon, **b)** bowtie2 filtering for
1488 uniquely mapping paired reads with MAPQ >0, and **c)** bowtie2 filtering for uniquely mapping
1489 paired reads with a MAPQ > 20. **d)** The per patient approach was applied. For each patient, the
1490 paired *ex vivo* and *in vitro* samples were analysed. The assembled *var* transcripts (at least 1500nt
1491 and containing 3 significantly annotated *var* domains) across all the generations for a patient were
1492 combined into a reference, redundancy was removed using cd-hit (at sequence identity = 99%),
1493 and expression was quantified using Salmon. Pie charts show the *var* expression profile with the
1494 relative size of each slice representing the relative percentage of total *var* gene expression of each
1495 *var* transcript. Different colours represent different assembled *var* transcripts with the same
1496 colour code used across a-d.
1497
1498 **Figure 4: Rank *var* gene expression analysis.** For each patient, the paired *ex vivo* (n=13) and *in*
1499 *vitro* samples (generation 1: n=13, generation 2: n=10, generation 3: n=1) were analysed. The
1500 assembled *var* transcripts with at least 1500nt and containing 3 significantly annotated *var*
1501 domains across all the generations for a patient were combined into a reference, redundancy was
1502 removed using cd-hit (at sequence identity = 99%), and expression was quantified using Salmon.
1503 *Var* transcript expression was ranked. Plots show the top 10 *var* gene expression rankings for each
1504 patient and their *ex vivo* and short-term *in vitro* cultured parasite samples. Group A *var* transcripts
1505 (red), group B or C *var* transcripts (blue), group E *var* transcripts (green) and transcripts of
1506 unknown *var* group (orange).
1507
1508 **Figure 4 – Figure supplement 1: Rank *var* gene expression analysis.** For each patient, the paired
1509 *ex vivo* (n=13) and *in vitro* samples (generation 1: n=13, generation 2: n=10, generation 3: n=1)
1510 were analysed. The assembled *var* transcripts with at least 1500nt and containing 3 significantly
1511 annotated *var* domains across all the generations for a patient were combined into a reference,
1512 redundancy was removed using cd-hit (at sequence identity = 99%), and expression was quantified
1513 using Salmon. *Var* transcript expression was ranked. Plots show the top 10 *var* gene expression
1514 rankings for each patient and their *ex vivo* and short-term *in vitro* cultured parasite samples. Group

1515 B/C *var* transcript (blue), group A *var* transcript (red), group E *var* transcript (green) and unknown
1516 (orange).
1517
1518 **Figure 4 – Figure supplement 2: *Var* expression homogeneity (VEH).** α diversity curves were
1519 determined on a per patient basis using Equation 1. *D* is calculated for *q* in the range 0 to 3 with a
1520 step increase of 0.1 and *p* in this analysis represented the proportion of *var* gene expression
1521 dedicated to *var* transcript *k*. *q* determined how much weight is given to rare vs abundant *var*
1522 transcripts. The smaller the *q* value, the less weight was given to the more abundant *var* transcript.

Equation 1 $$D_{(q)} = (\sum_{k=1}^{K} p_k^q)^{\frac{1}{1-q}}$$

1523
1524
1525 **Figure 4 – Figure supplement 3: *Var2csa* expression through short-term *in vitro* cultivation.** All
1526 significantly annotated *var2csa* transcripts had their $\log_2$ Salmon normalised count summed.
1527 Significant differences in expression levels between generations were assessed using a paired
1528 Wilcoxon test and adjusted using the Benjamini-Hochberg method. Numbers above the boxplots
1529 represent significant differences observed (FDR <=0.05). Coloured lines connect paired patient
1530 samples through the generations.
1531
1532 **Figure 5: *Var* domain transcriptome analysis through short-term *in vitro* culture.** *Var*
1533 transcripts for paired *ex vivo* (n=13), generation 1 (n=13), generation 2 (n=10) and generation 3
1534 (n=1) were *de novo* assembled using the whole transcript approach. *Var* transcripts were filtered
1535 for those >= 1500nt in length and containing at least 3 significantly annotated *var* domains.
1536 Transcripts were annotated using HMM models built on the Rask et al., 2010 dataset (Rask *et al.*,
1537 2010). When annotating the whole transcript, the most significant alignment was taken as the
1538 best annotation for each region of the assembled transcript (e-value cut off 1e-5). Multiple
1539 annotations were allowed on the transcript if they were not overlapping, determined using cath-
1540 resolve-hits. *Var* domain expression was quantified using FeatureCounts and the domain counts
1541 aggregated **a)** PCA plot of $\log_2$ normalized read counts (adjusted for life cycle stage, derived from
1542 the mixture model approach). Points are coloured by their generation (*ex vivo*; purple,
1543 generation 1; red, generation 2; green and generation 3; blue) and labelled by their patient
1544 identity **b)** Volcano plot showing extent and significance of up- or down-regulation of *var* domain
1545 expression in *ex vivo* (n=13) compared with paired generation 1 cultured parasites (n=13) (red
1546 and blue, $P < 0.05$ after Benjamini-Hochberg adjustment for FDR; red and green, absolute
1547 $\log_2$ fold change $\log_2$FC in expression >= 2). Domains with a $\log_2$FC > = 2 represent those
1548 upregulated in generation 1 parasites. Domains with a $\log_2$FC <= -2 represent those
1549 downregulated in generation 1 parasites. **c)** Volcano plot showing extent and significance of up-
1550 or down-regulation of *var* domain expression in *ex vivo* (n=10) compared with paired generation
1551 2 cultured parasites (n=10) (green, absolute $\log_2$ fold change $\log_2$FC in expression >= 2). Domains
1552 with a $\log_2$FC >= 2 represent those upregulated in generation 2 parasites. Domains with a $\log_2$FC
1553 <= -2 represent those downregulated in generation 2 parasites. Differential expression analysis
1554 was performed using DESeq2 (adjusted for life cycle stage, derived from the mixture model
1555 approach).
1556
1557 **Figure 6. *Var* group expression analysis through short-term *in vitro* culture.** The DBLα domain
1558 sequence for each transcript was determined and for each patient a reference of all assembled
1559 DBLα domains combined. Group A *var* genes possess DBLα1 domains, some group B encode
1560 DBLα2 domains and groups B and C encode DBLα0 domains. Domains were grouped by type and
1561 their expression summed. The relevant sample's non-core reads were mapped to this using

1562 Salmon and DBLα expression quantified. DESeq2 normalisation was performed, with patient
1563 identity and life cycle stage proportions included as covariates. A similar approach was repeated
1564 for NTS domains. Group A *var* genes encode NTSA compared to group B and C *var* genes which
1565 encode NTSB. Boxplots show $\log_2$ normalised Salmon read counts for **a)** group A *var* gene
1566 expression through cultured generations assessed using the DBLα domain sequences, **b)** group B
1567 and C *var* gene expression through cultured generations assessed using the DBLα domain
1568 sequences, **c)** group A *var* gene expression through cultured generations assessed using the NTS
1569 domain sequences, and **d)** group B and C *var* gene expression through cultured generations
1570 assessed using the NTS domain sequences. Different coloured lines connect paired patient
1571 samples through the generations: *ex vivo* (n=13), generation 1 (n=13), generation 2 (n=10) and
1572 generation 3 (n=1). Axis shows different scaling.
1573
1574 **Figure 6 – Figure supplement 1: Total *var* gene expression through short-term *in vitro* culture.**
1575 Boxplots show data from paired *ex vivo* (n=13), generation 1 (n=13), generation 2 (n=10) and
1576 generation 3 (n=1) parasite samples. The RNA-sequencing reads were blastn (with the short-blastn
1577 option on and significance = 1e-10) against the LARSFADIG nucleotide sequences (142 unique
1578 LARSFADIG sequences) to identify reads containing the LARSFADIG motifs. Once the reads
1579 containing the LARSFADIG motifs had been identified, they were used to assemble the LARSFADIG
1580 motif. rnaSPAdes and Trinity were used to assemble the LARSFADIG motif. The sequencing reads
1581 were mapped back against the assemblies using bwa mem, and coverage over the middle of the
1582 motif (S) determined. These values were divided by the number of reads mapped to the *var* exon
1583 1 database and the 3D7 genome (which had *var* genes removed) to represent the proportion of
1584 total gene expression dedicated to *var* gene expression (similar to an RPKM). % total gene
1585 expression dedicated to *var* gene expression quantified using LARSFADIG motifs assembled **a)** by
1586 Trinity, **b)** by rnaSPAdes, or **c)** # assembled LARSFADIG motifs, assembled using rnaSPAdes.
1587
1588 **Figure 6 – Figure supplement 2: Verification of RNA-sequencing results using DBLα-tag**
1589 **sequencing.** Amplified DBLα-tag sequences were blasted against the ~2400 genomes on varDB to
1590 obtain subclassification into DBLα0/1/2 and prediction of adjacent head structure N-terminal
1591 segment and cysteine-rich interdomain region domains and their related binding phenotype.
1592 Proportion of each NTS and DBLα subclass were calculated, and group A and group B and C
1593 expression proportions determined. Group A *var* genes encode NTSA and DBLα1 and groups B and
1594 C encode NTSB and DBLα0. Group B *var* genes encode DBLα2. **a)** Group A transcript proportions
1595 determined by the DBLα1 domain. **b)** Group B and C transcript proportions determined by DBLα0
1596 and DBLα2 domains. **c)** Group A transcript proportions determined by the NTSA domain. **d)** Group
1597 B and C transcript proportions determined by the NTSB domain.
1598
1599 **Figure 7: Core gene transcriptome analysis of *ex vivo* and short-term *in vitro* cultured samples.**
1600 Core gene expression was assessed for paired *ex vivo* (n=13), generation 1 (n=13), generation 2
1601 (n=10) and generation 3 (n=1) parasite samples. Subread align was used, as in the original analysis,
1602 to align the reads to the human genome and *P. falciparum* 3D7 genome, *with var, rif, stevor, surf*
1603 and *rRNA* genes removed. HTSeq count was used to quantify gene counts. **a)** PCA plot of $\log_2$
1604 normalized read counts. Points are coloured by their generation (*ex vivo*: purple, generation 1: red,
1605 generation 2: green, and generation 3: blue) and labelled by their patient identity. **b)** Volcano plot
1606 showing extent and significance of up- or down-regulation of core gene expression in *ex vivo* (n=13)
1607 compared with paired generation 1 cultured parasites (n=13) and **c)** in *ex vivo* (n=10) compared
1608 with paired generation 2 cultured parasites (n=10). Dots in red and blue represent those genes
1609 with $P < 0.05$ after Benjamini-Hochberg adjustment for FDR, red and green dots label genes with
1610 absolute $\log_2$ fold change $\log_2$FC in expression >= 2). Accordingly, genes with a $\log_2$FC > = 2

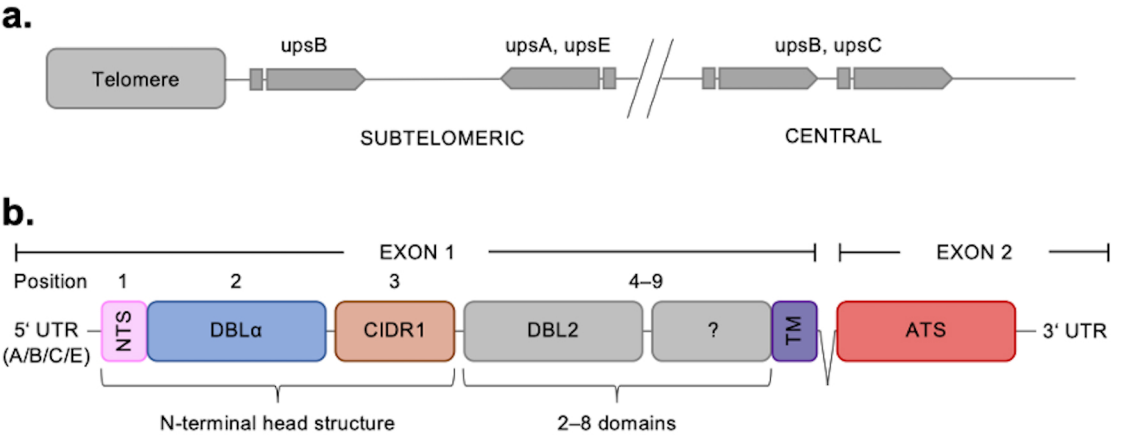
represent those upregulated in generation 1 parasites and genes with a log_2FC <= -2 represent those downregulated in generation 1 parasites. Normalized read counts of the core gene analysis were adjusted for life cycle stage, derived from the mixture model approach.

Figure 7 – Figure supplement 1: Top 20 enriched pathways in the core genes found to be significantly upregulated in generation 1 parasite samples (n=13) compared to *ex vivo* parasite samples (n=13). The 902 genes found significantly upregulated in generation 1 parasite samples underwent pathway analysis. Gene ontology and KEGG analysis was performed using ShinyGo and significant terms were defined by having a Bonferroni corrected p-value < 0.05. **a)** Network of enriched pathways. Two pathways (nodes) are connected if they share 20% or more genes. Darker nodes are more significantly enriched gene sets. Bigger nodes represent larger gene sets. Thicker edges represent more overlapped genes. **b)** Fold enrichment, significance level and the number of genes found in the pathways.

Figure 7 – Figure supplement 2: Verification of core gene expression analysis excluding schizont and gametocyte stage parasite samples. Core gene expression was assessed for paired *ex vivo* (n=6) and generation 1 (n=6) parasite samples where there were no schizont or gametocyte stage parasites, as determined using the mixture model approach. 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. HTSeq count was used to quantify gene counts. **a)** Volcano plot showing extent and significance of up- or down-regulation of core gene expression in *ex vivo* compared with paired generation 1 cultured parasites (red and blue: $P < 0.05$ after Benjamini-Hochberg adjustment for FDR, red and green: absolute log_2 fold change log_2FC in expression >= 2). Genes with a log_2FC > = 2 represent those upregulated in generation 1 parasites. Genes with a log_2FC <= -2 represent those downregulated in generation 1 parasites. **b)** Spearman's rank correlation analysis of the log_2FC values for the *ex vivo* vs generation 1 core gene differential expression analysis using all paired *ex vivo* (n=13) and generation 1 (n=13) samples vs *ex vivo* (n=9) and generation 1 (n=9) samples containing no schizont or gametocyte stage parasites.

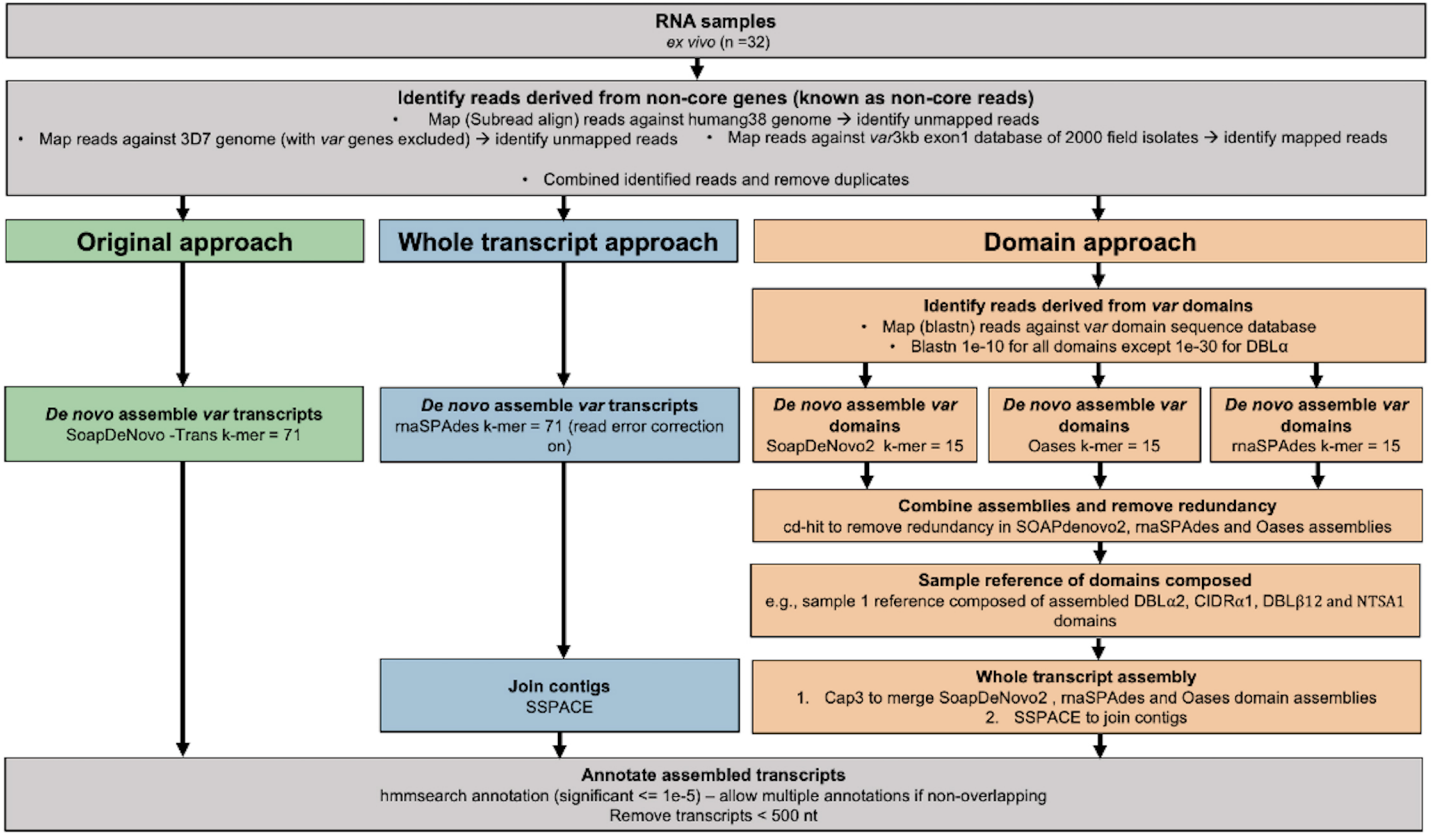
Figure 7 – Figure supplement 3: Short-term *in vitro* cultured parasites as surrogates for assessing *the in vivo* core gene transcriptome. Core gene expression was assessed for paired *ex vivo* (n=13) and generation 1 (n=13) parasite samples. Subread align was used, as in the original analysis, to align the reads to the human genome and *P. falciparum* 3D7 genome, *with var, rif, stevor, surf* and *rRNA* genes removed. HTSeq count was used to quantify gene counts (Anders et al., 2015). **a)** Volcano plot showing extent and significance of up- or down-regulation of core gene expression in malaria naïve patient *ex vivo* samples (n=6) compared with previously exposed malaria patient *ex vivo* samples (n=7) (red and blue: $P < 0.05$ after Benjamini-Hochberg adjustment for FDR; red and green: absolute log_2 fold change log_2FC in expression >= 2). Genes with a log_2FC > = 2 represent those upregulated in malaria naïve patients. Genes with a log_2FC <= -2 represent those downregulated in malaria naïve patients. **b)** Volcano plot showing extent and significance of up- or down-regulation of core gene expression in malaria naïve patient samples that underwent one cycle of *in vitro* cultivation (n=6) compared with previously exposed malaria patient samples that underwent one cycle of *in vitro* cultivation (n=7) (red and blue, $P < 0.05$ after Benjamini-Hochberg adjustment for FDR; red and green, absolute log_2 fold change log_2FC in expression >= 2). Genes with a log_2FC > = 2 represent those upregulated in malaria naïve patients. Genes with a log_2FC <= -2 represent those downregulated in malaria naïve patients. Differential expression analysis was performed using DESeq2 (adjusted for life cycle stage, derived from the mixture model approach).

1660 **Supplementary file 1: Output of the linear model to quantify total *var* gene expression during**
1661 **the transition of parasites into *in vitro* culture.** *Var* gene expression as a proportion of total gene
1662 expression was used as a response variable, generation and life cycle stage as independent
1663 variables and patient information as a random effect. Only paired samples of *ex vivo* parasites and
1664 generation 1 parasites were used.
1665
1666 **Supplementary file 2: Varia results output.** For each patient sample, the corresponding paired
1667 cluster is labelled (e.g., #6_1st_invitro_cluster_154=cluster 90ev shows cluster 154 of patient #6
1668 generation 1 sample corresponds to cluster 90 of patient #6 *ex vivo* sample. Identical cluster
1669 sequences across paired samples were defined as having % sequence ID = > 99%.
1670
1671 **Supplementary file3: Differentially expressed core genes in the *ex vivo* vs generation 1 paired**
1672 **analysis.** Log2FoldChange represents the log2 fold change for the *ex vivo* and generation 1
1673 analysis. Values >0 represent genes up-regulated in generation 1 samples and values <0 represent
1674 genes down-regulated in generation 1 samples. pvalue and padj represent the p-value and
1675 adjusted pvalue (Benjamini Hochberg) respectively.
1676
1677 **Supplementary file 4: MSP1 genotyping results from *ex vivo* and *in vitro*-adapted *P. falciparum***
1678 **isolates.**
1679
1680 **Supplementary file 5: Quality of the RNA samples analyzed in this study.** To characterize the
1681 overall RNA quality prior to library synthesis, the Bioanalyzer automated RNA electrophoresis
1682 system was used to visualize the samples and calculate the RIN values. The measurement of the
1683 RIN value for samples from mixed species (*H. sapiens*, *P. falciparum*) is not very meaningful, as the
1684 RIN increases the higher the proportion of a single species. This can be observed by the increase
1685 in the RIN value during *in vitro* cultivation of the parasites, as the parasite RNA content increases
1686 over time. Of the four rRNA peaks visible in particular in the *ex vivo* samples, the inner peaks
1687 represent the 18S and 28S rRNA of *P. falciparum*, the outer peaks are of human origin.

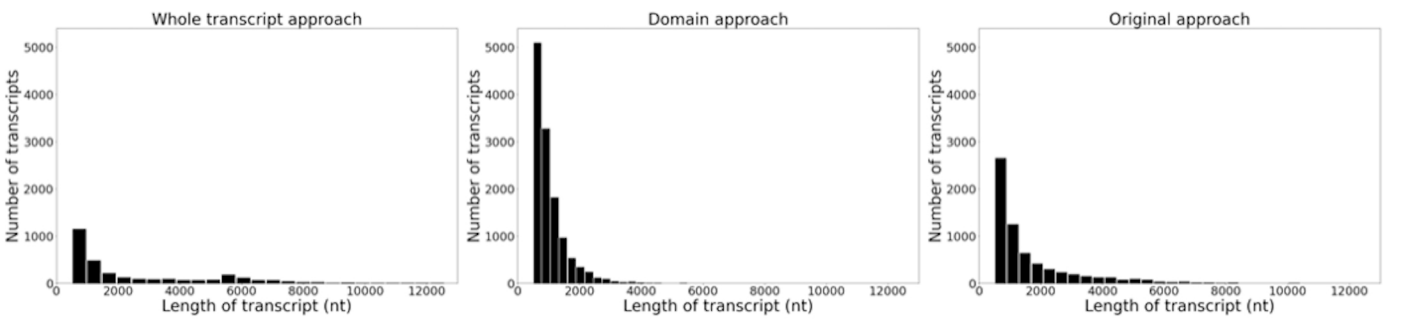


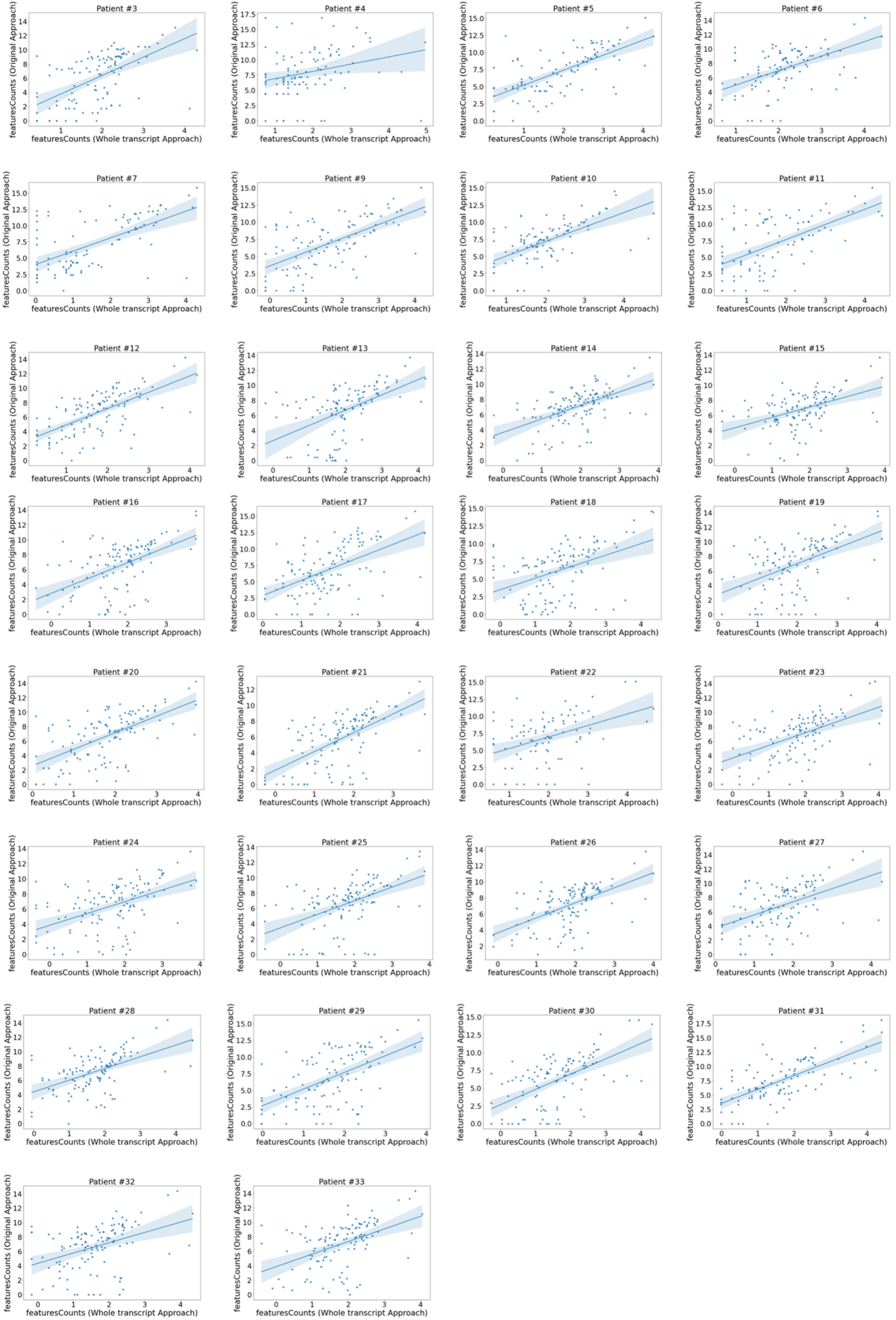
c.

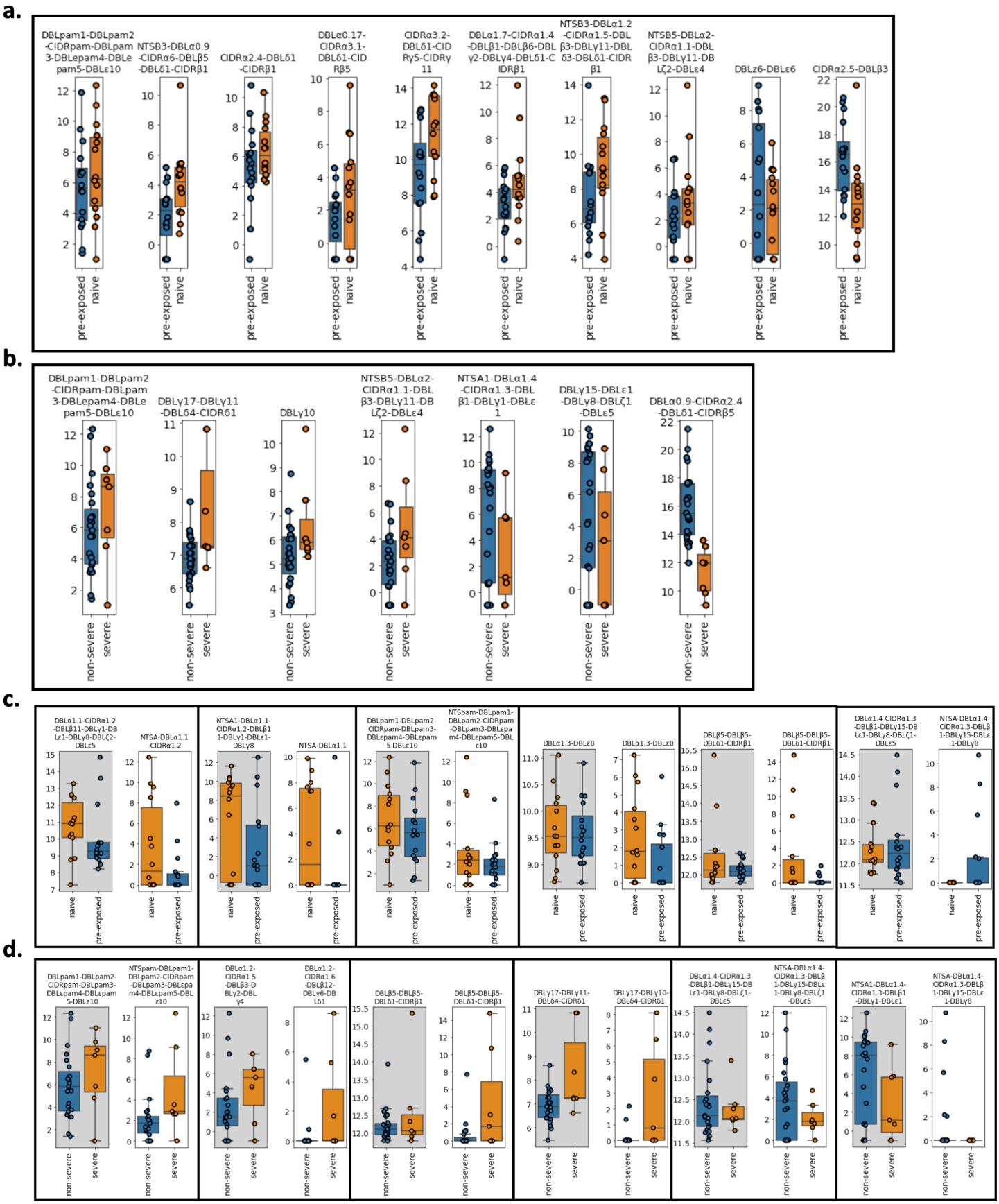
	Binding phenotype	*Var* group	Position 1	Position 2	Position 3	Position 4–9
‚Regular' *var* genes	Unknown	A	NTSA	DBLα1	CIDRδ/β/γ	DBLβ, DBLγ, DBLδ, CIDRδ/β/γ, DBLζ, DBLε
	EPCR	A	NTSA	DBLα1	CIDRα1.4–7	
	EPCR	B	NTSB	DBLα2	CIDRα1.1/8	
	CD36	B & C	NTSB	DBLα0	CIDRα2–6	
Conserved *var* genes	CSA	E, *var2csa*	NTSpam	DBLpam1	DBLpam2	CIDRpam, DBLpam3, DBLεpam4, DBLεpam5, DBLεpam6
	Unknown	A, *var1-3D7*	NTSA	DBLα1.4	CIDRα1.3	DBLβ1, DBLγ15, DBLε1, DBLγ8, DBLζ1, DBLε5
	Unknown	A, *var1-IT*	NTSA	DBLα1.1	CIDRα1.2	DBLβ11, DBLγ1, DBLε1, DBLγ8, DBLζ2, DBLε5
	Unknown	A, *var3*	NTSA	DBLα1.3	–	DBLε8

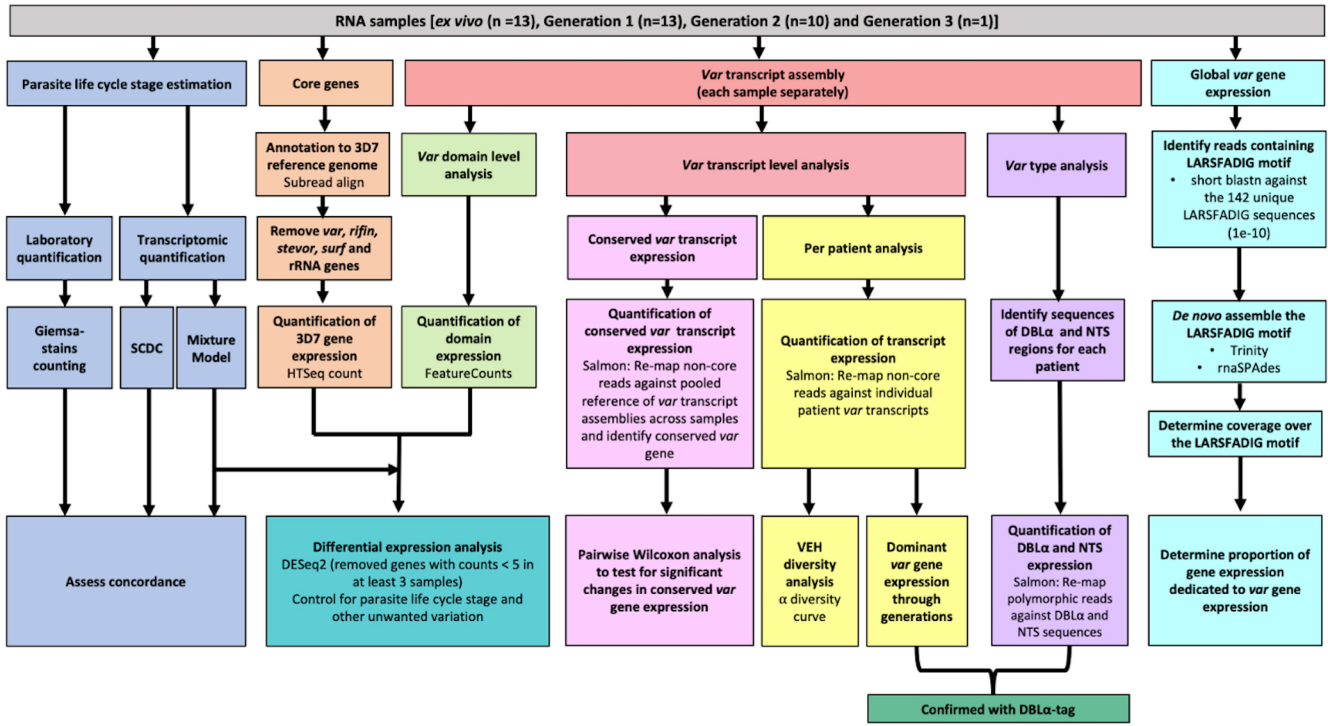

RNA samples
ex vivo (n =32)
Identify reads derived from non-core genes (known as non-core reads)
• Map (Subread align) reads against humang38 genome → identify unmapped reads
• Map reads against 3D7 genome (with var genes excluded) → identify unmapped reads
• Map reads against var3kb exon1 database of 2000 field isolates → identify mapped reads
• Combined identified reads and remove duplicates
Original approach
Whole transcript approach
Domain approach
Identify reads derived from var domains
• Map (blastn) reads against var domain sequence database
• Blastn 1e-10 for all domains except 1e-30 for DBLα
De novo assemble var transcripts
SoapDeNovo -Trans k-mer = 71
De novo assemble var transcripts
rnaSPAdes k-mer = 71 (read error correction on)
De novo assemble var domains
SoapDeNovo2 k-mer = 15
De novo assemble var domains
Oases k-mer = 15
De novo assemble var domains
rnaSPAdes k-mer = 15
Combine assemblies and remove redundancy
cd-hit to remove redundancy in SOAPdenovo2, rnaSPAdes and Oases assemblies
Sample reference of domains composed
e.g., sample 1 reference composed of assembled DBLα2, CIDRα1, DBLβ12 and NTSA1 domains
Join contigs
SSPACE
Whole transcript assembly
1. Cap3 to merge SoapDeNovo2 , rnaSPAdes and Oases domain assemblies
2. SSPACE to join contigs
Annotate assembled transcripts
hmmsearch annotation (significant <= 1e-5) – allow multiple annotations if non-overlapping
Remove transcripts < 500 nt

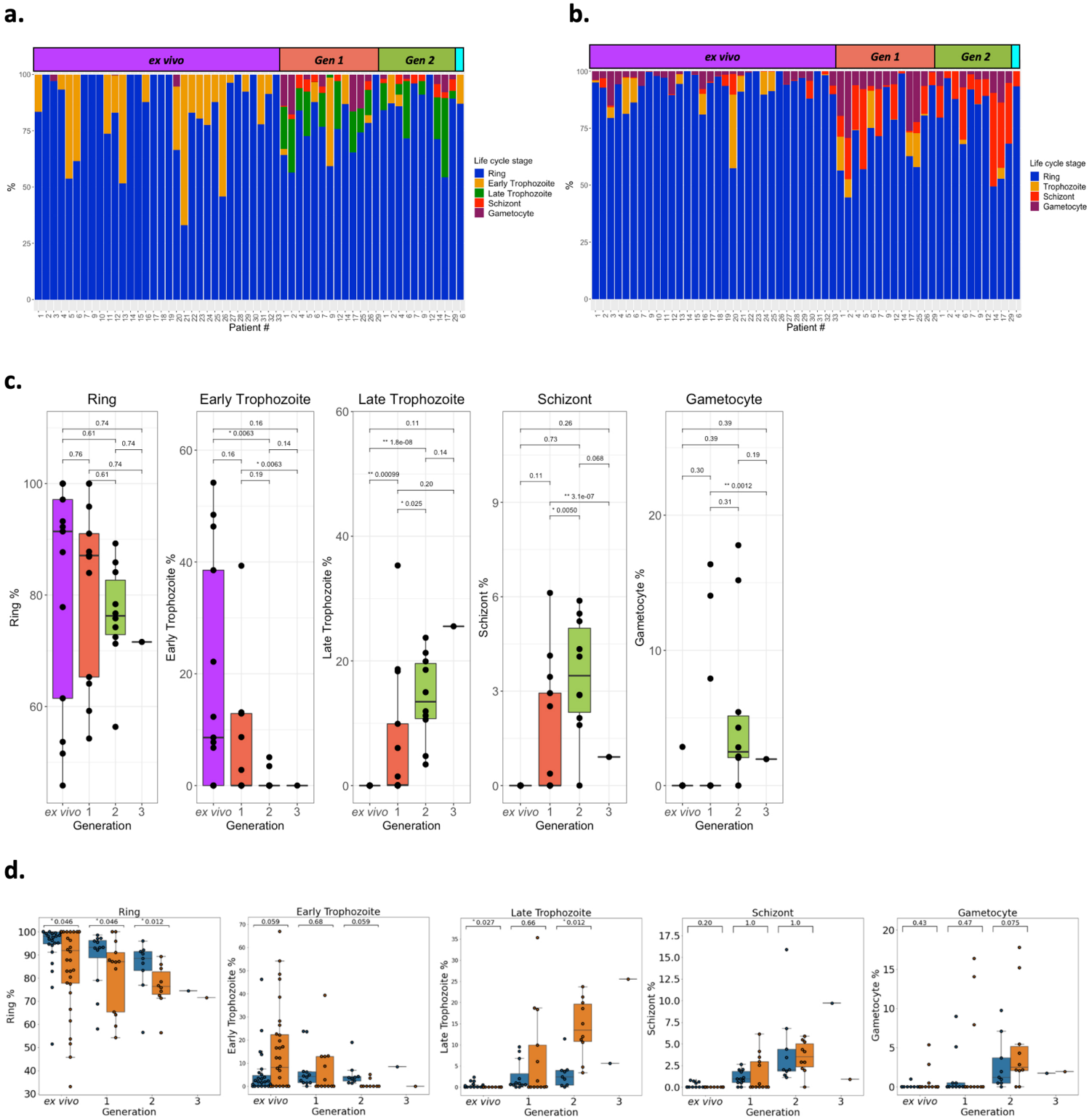
a.
Alignment score
Time (hpi)
b.
Accuracy score
Time (hpi)
c.
No. contigs
Time (hpi)
d.
Alignment score
Time (hpi)
e.
Alignment score
Time (hpi)
f.
Whole transcript approach
Domain approach
Original approach
48 hours
44 hours
40 hours
32 hours
24 hours
16 hours
8 hours
0 hours
reference transcript (nt)
g.
Whole transcript approach
Domain approach
Original approach
reference transcript (nt)
h.
Whole transcript approach
Domain approach
Original approach
reference transcript (nt)

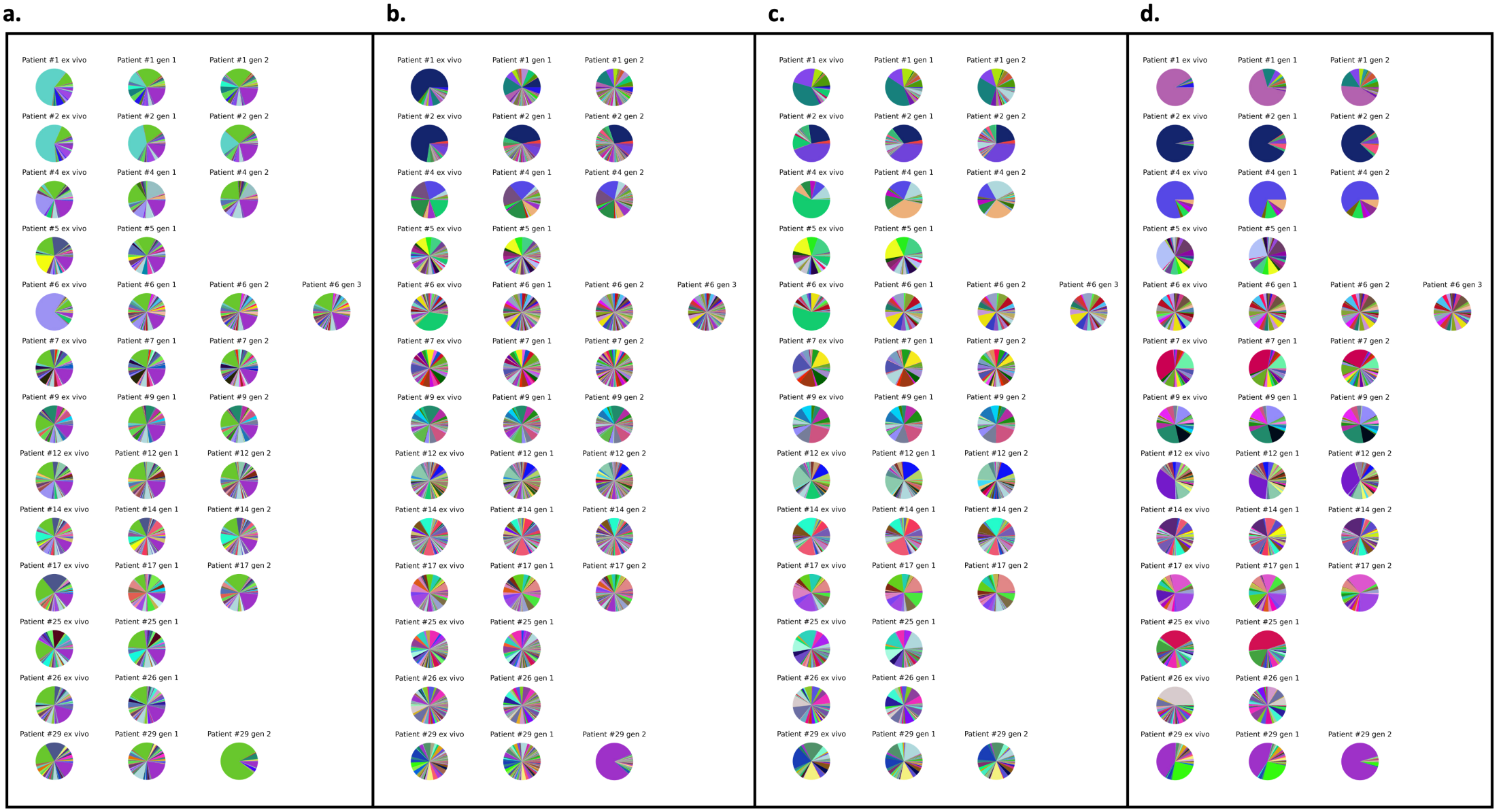

Whole transcript approach
Domain approach
Original approach
Number of transcripts
Length of transcript (nt)
0
1000
2000
3000
4000
5000
0
2000
4000
6000
8000
10000
12000


Patient #3
Patient #4
Patient #5
Patient #6
Patient #7
Patient #9
Patient #10
Patient #11
Patient #12
Patient #13
Patient #14
Patient #15
Patient #16
Patient #17
Patient #18
Patient #19
Patient #20
Patient #21
Patient #22
Patient #23
Patient #24
Patient #25
Patient #26
Patient #27
Patient #28
Patient #29
Patient #30
Patient #31
Patient #32
Patient #33
featuresCounts (Original Approach)
featuresCounts (Whole transcript Approach)

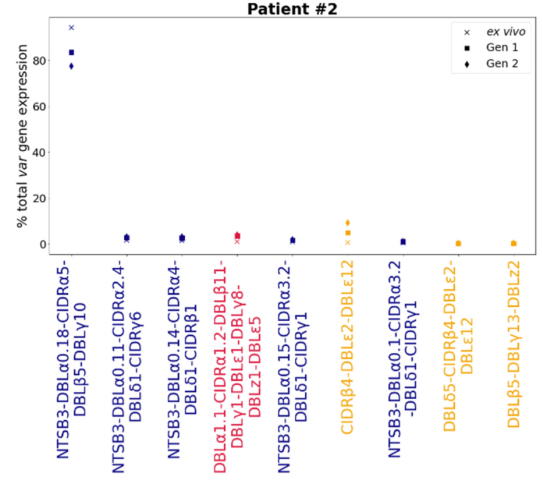

a.
DBLpam1-DBLpam2-CIDRpam-DBLpam3-DBLepam4-DBLepam5-DBLε10
NTSB3-DBLα0.9-CIDRα6-DBLβ5-DBLδ1-CIDRβ1
CIDRα2.4-DBLδ1-CIDRβ1
DBLα0.17-CIDRα3.1-DBLδ1-CIDRβ5
CIDRα3.2-DBLδ1-CIDRγ5-CIDRγ11
DBLα1.7-CIDRα1.4-DBLβ1-DBLβ6-DBLγ2-DBLγ4-DBLδ1-CIDRβ1
NTSB3-DBLα1.2-CIDRα1.5-DBLβ3-DBLγ11-DBLδ3-DBLδ1-CIDRβ1
NTSB5-DBLα2-CIDRα1.1-DBLβ3-DBLγ11-DBLζ2-DBLε4
DBLz6-DBLε6
CIDRα2.5-DBLβ3
pre-exposed
naive
b.
DBLpam1-DBLpam2-CIDRpam-DBLpam3-DBLepam4-DBLepam5-DBLε10
DBLγ17-DBLγ11-DBLδ4-CIDRδ1
DBLγ10
NTSB5-DBLα2-CIDRα1.1-DBLβ3-DBLγ11-DBLζ2-DBLε4
NTSA1-DBLα1.4-CIDRα1.3-DBLβ1-DBLγ1-DBLε1
DBLγ15-DBLε1-DBLγ8-DBLζ1-DBLε5
DBLα0.9-CIDRα2.4-DBLδ1-CIDRβ5
non-severe
severe
c.
DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLζ2-DBLε5
NTSA-DBLα1.1-CIDRα1.2
NTSA1-DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8
NTSA-DBLα1.1
DBLpam1-DBLpam2-CIDRpam-DBLpam3-DBLεpam4-DBLεpam5-DBLε10
NTSpam-DBLpam1-DBLpam2-CIDRpam-DBLpam3-DBLεpam4-DBLεpam5-DBLε10
DBLα1.3-DBLε8
DBLα1.3-DBLε8
DBLβ5-DBLβ5-DBLδ1-CIDRβ1
DBLβ5-DBLβ5-DBLδ1-CIDRβ1
DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLγ8-DBLζ1-DBLε5
NTSA-DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLγ8
naive
pre-exposed
d.
DBLpam1-DBLpam2-CIDRpam-DBLpam3-DBLepam4-DBLepam5-DBLε10
NTSpam-DBLpam1-DBLpam2-CIDRpam-DBLpam3-DBLεpam4-DBLεpam5-DBLε10
DBLα1.2-CIDRα1.5-DBLβ3-DBLγ2-DBLγ4
DBLα1.2-CIDRα1.6-DBLβ12-DBLγ6-DBLδ1
DBLβ5-DBLβ5-DBLδ1-CIDRβ1
DBLβ5-DBLβ5-DBLδ1-CIDRβ1
DBLγ17-DBLγ11-DBLδ4-CIDRδ1
DBLγ17-DBLγ10-DBLδ4-CIDRδ1
DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLζ1-DBLε5
NTSA-DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLγ8-DBLζ1-DBLε5
NTSA1-DBLα1.4-CIDRα1.3-DBLβ1-DBLγ1-DBLε1
NTSA-DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLγ8
non-severe
severe

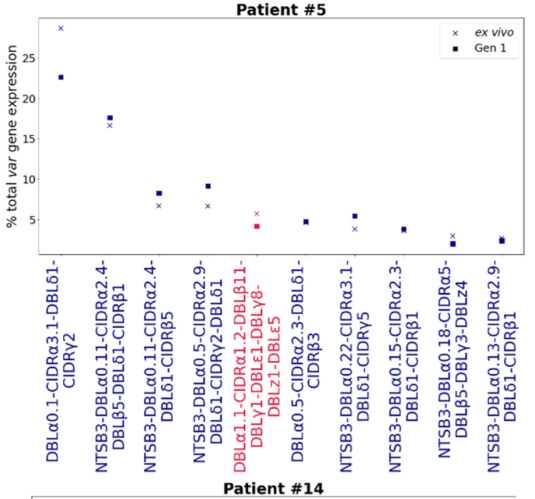

RNA samples [ex vivo (n =13), Generation 1 (n=13), Generation 2 (n=10) and Generation 3 (n=1)]
Parasite life cycle stage estimation
Core genes
Var transcript assembly (each sample separately)
Global var gene expression
Laboratory quantification
Transcriptomic quantification
Giemsa-stains counting
SCDC
Mixture Model
Annotation to 3D7 reference genome
Subread align
Remove var, rifin, stevor, surf and rRNA genes
Quantification of 3D7 gene expression
HTSeq count
Var domain level analysis
Quantification of domain expression
FeatureCounts
Var transcript level analysis
Conserved var transcript expression
Per patient analysis
Quantification of conserved var transcript expression
Salmon: Re-map non-core reads against pooled reference of var transcript assemblies across samples and identify conserved var gene
Quantification of transcript expression
Salmon: Re-map non-core reads against individual patient var transcripts
Var type analysis
Identify sequences of DBLα and NTS regions for each patient
Identify reads containing LARSFADIG motif
• short blastn against the 142 unique LARSFADIG sequences (1e-10)
De novo assemble the LARSFADIG motif
• Trinity
• rnaSPAdes
Determine coverage over the LARSFADIG motif
Assess concordance
Differential expression analysis
DESeq2 (removed genes with counts < 5 in at least 3 samples)
Control for parasite life cycle stage and other unwanted variation
Pairwise Wilcoxon analysis to test for significant changes in conserved var gene expression
VEH diversity analysis
α diversity curve
Dominant var gene expression through generations
Quantification of DBLα and NTS expression
Salmon: Re-map polymorphic reads against DBLα and NTS sequences
Determine proportion of gene expression dedicated to var gene expression
Confirmed with DBLα-tag

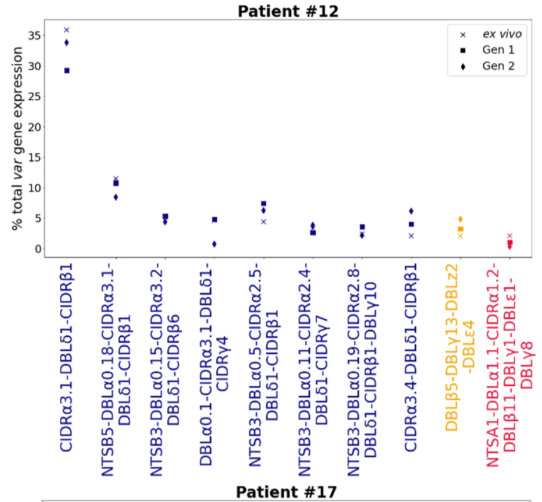

a.
ex vivo
Gen 1
Gen 2
%
Patient #
Life cycle stage
Ring
Early Trophozoite
Late Trophozoite
Schizont
Gametocyte
b.
ex vivo
Gen 1
Gen 2
%
Patient #
Life cycle stage
Ring
Trophozoite
Schizont
Gametocyte
c.
Ring
Early Trophozoite
Late Trophozoite
Schizont
Gametocyte
Ring %
Early Trophozoite %
Late Trophozoite %
Schizont %
Gametocyte %
ex vivo
Generation
d.
Ring
Early Trophozoite
Late Trophozoite
Schizont
Gametocyte
Ring %
Early Trophozoite %
Late Trophozoite %
Schizont %
Gametocyte %
ex vivo
Generation

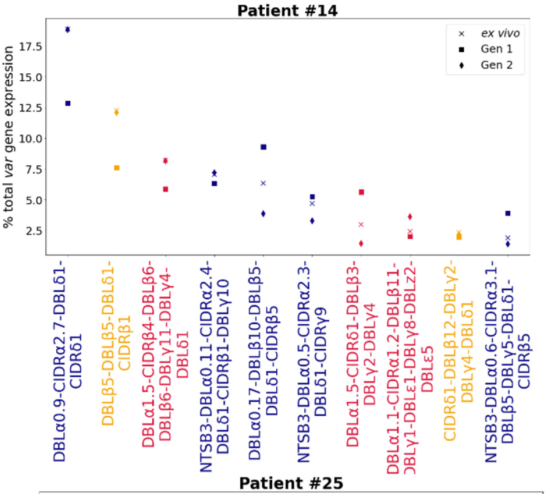

a.
b.
c.
d.
Patient #1 ex vivo
Patient #1 gen 1
Patient #1 gen 2
Patient #2 ex vivo
Patient #2 gen 1
Patient #2 gen 2
Patient #4 ex vivo
Patient #4 gen 1
Patient #4 gen 2
Patient #5 ex vivo
Patient #5 gen 1
Patient #6 ex vivo
Patient #6 gen 1
Patient #6 gen 2
Patient #6 gen 3
Patient #7 ex vivo
Patient #7 gen 1
Patient #7 gen 2
Patient #9 ex vivo
Patient #9 gen 1
Patient #9 gen 2
Patient #12 ex vivo
Patient #12 gen 1
Patient #12 gen 2
Patient #14 ex vivo
Patient #14 gen 1
Patient #14 gen 2
Patient #17 ex vivo
Patient #17 gen 1
Patient #17 gen 2
Patient #25 ex vivo
Patient #25 gen 1
Patient #26 ex vivo
Patient #26 gen 1
Patient #29 ex vivo
Patient #29 gen 1
Patient #29 gen 2

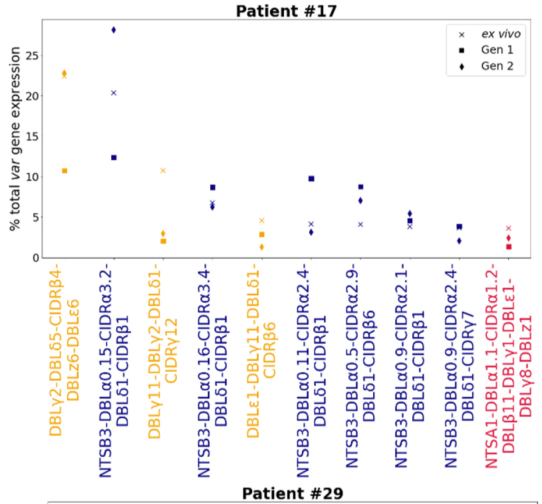
Patient #1
% total var gene expression
ex vivo
Gen 1
Gen 2
DBLβ5-DBLβ5-DBLδ1-CIDRβ1
CIDRα2.10-DBLβ5-DBLβ5-DBLδ1-CIDRβ1
DBLα1.4-CIDRα1.1-DBLβ12-DBLγ2-DBLγ4
DBLδ1-DBLβ5-DBLβ5-DBLδ69-DBLδ1-CIDRβ5
NTSB5-DBLα2-CIDRα1.1-DBLβ12-DBLγ5-DBLδ1-CIDRβ6
DBLγ5-DBLδ1-CIDRβ6
DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLz1-DBLε5
CIDRα2.1-DBLβ5-DBLγ5-DBLδ1-CIDRβ5
NTSB3-DBLα0.12-CIDRα2.1-DBLδ1-CIDRβ1-DBLδ1
DBLβ6-DBLγ6-DBLδ5-DBLδ1-CIDRβ5
Patient #4
% total var gene expression
ex vivo
Gen 1
Gen 2
DBLα0.9-CIDRα2.7-DBLδ5-CIDRβ3-DBLβ9-DBLγ11
NTSB3-DBLα0.22-CIDRα3.1-DBLδ1-CIDRβ5
NTSB3-DBLα0.15-CIDRα4-DBLδ1-CIDRβ1
DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLγ8-DBLz2-DBLε5
DBLα0.9-CIDRα2.7-DBLδ1-CIDRγ4
DBLα0.5-CIDRα2.8-DBLγ11-DBLδ1-CIDRγ2
DBLεpαm4-DBLεpαm5-DBLε10
Patient #6
% total var gene expression
ex vivo
Gen 1
Gen 2
Gen 3
NTSB3-DBLα0.13-CIDRα2.3-DBLδ1-CIDRβ1
NTSB3-DBLα0.15-CIDRα3.2-DBLδ1-CIDRγ2-DBLε4
DBLα0.9-CIDRα2.2-DBLδ1-CIDRβ1
DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLz1-DBLε5
NTSB3-DBLα0.16-CIDRα3.4-DBLδ1-CIDRγ6
NTSB5-DBLα0.5-CIDRα2.10-DBLβ8-DBLδ1-CIDRβ1
NTSB3-DBLα0.5-CIDRα3.2-DBLδ1-CIDRγ2-DBLε4
NTSB7-DBLα0.16-CIDRα3.4-DBLδ1-CIDRβ5
NTSB3-DBLα0.9-CIDRα2.2-DBLδ1-CIDRβ1
DBLα0.18-CIDRα6-DBLβ5-DBLγ10
Patient #7
% total var gene expression
ex vivo
Gen 1
Gen 2
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRβ1
DBLα1.4-CIDRα1.3-DBLβ1-DBLγ15-DBLε1-DBLγ8-DBLz1-DBLε5
NTSB3-DBLα0.5-CIDRα2.6-DBLδ1-CIDRβ1
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRβ5
DBLα0.11-CIDRα2.4-DBLδ1-CIDRγ1
NTSB3-DBLα0.22-CIDRα3.1-DBLδ1-CIDRβ1
DBLα0.5-CIDRα2.6-DBLδ1-CIDRβ1
NTSB3-DBLα0.6-CIDRα3.2-DBLβ5-DBLγ11-DBLz4
NTSB3-DBLα0.9-CIDRα6-DBLβ5-DBLγ11-DBLz1-DBLε9
NTSB3-DBLα0.9-CIDRα2.4-DBLδ1-CIDRβ1
Patient #9
% total var gene expression
ex vivo
Gen 1
Gen 2
NTSB3-DBLα0.16-CIDRα3.4-DBLδ4-DBLδ1-CIDRβ1
NTSB3-DBLα0.15-CIDRα2.5-DBLδ1-CIDRγ9
NTSA1-DBLα1.1-CIDRα1.2-DBLβ11
NTSB3-DBLα0.8-CIDRα3.4-DBLδ1-CIDRβ5
NTSB3-DBLα0.5-CIDRα2.10-DBLδ1-CIDRγ2
NTSB3-DBLα0.9-CIDRα2.4-DBLδ1-CIDRγ4
NTSB2-DBLα0.6-CIDRα3.2-DBLδ1-CIDRβ1
NTSB3-DBLα0.3-CIDRα6-DBLβ5-DBLβ5-DBLγ5-DBLδ1-CIDRβ1
NTSB3-DBLα0.5-CIDRα2.3-DBLδ1-CIDRβ1
NTSB7-DBLα0.9-CIDRα4-DBLδ1-CIDRβ1
Patient #26
% total var gene expression
ex vivo
Gen 1
DBLα1.7-CIDRα1.4-DBLβ3-DBLγ2-DBLγ6-DBLγ11-DBLδ1-CIDRβ1
NTSB3-DBLα2-CIDRα1.1-DBLβ12-DBLγ14-DBLγ12-DBLδ1-CIDRβ5
NTSB3-DBLα0.5-CIDRα2.3-DBLγ10-DBLδ1-CIDRβ1-DBLδ1
CIDRβ2-DBLγ7-DBLε11-DBLz2-DBLε6
NTSB4-DBLα0.6-CIDRα3.1-DBLδ1-CIDRβ5
NTSB3-DBLα0.16-CIDRα3.4-DBLδ1-CIDRγ5-DBLz6-DBLε9
NTSB3-DBLα0.22-CIDRα3.1-DBLδ1-CIDRβ1
DBLα1.6-CIDRγ3-DBLβ3-DBLγ12-DBLδ5-CIDRβ3-DBLβ7-DBLδ1
NTSB3-DBLα0.9-CIDRα2.1-DBLδ1-CIDRγ10
NTSB3-DBLα0.9-CIDRα2.2-DBLδ1-CIDRβ1

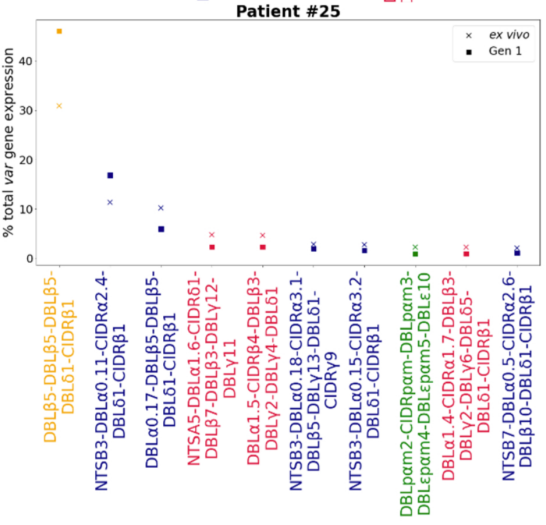

Patient #2
% total var gene expression
ex vivo
Gen 1
Gen 2
NTSB3-DBLα0.18-CIDRα5-DBLβ5-DBLγ10
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRγ6
NTSB3-DBLα0.14-CIDRα4-DBLδ1-CIDRβ1
DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLz1-DBLε5
NTSB3-DBLα0.15-CIDRα3.2-DBLδ1-CIDRγ1
CIDRβ4-DBLε2-DBLε12
NTSB3-DBLα0.1-CIDRα3.2-DBLδ1-CIDRγ1
DBLδ5-CIDRβ4-DBLε2-DBLε12
DBLβ5-DBLγ13-DBLz2

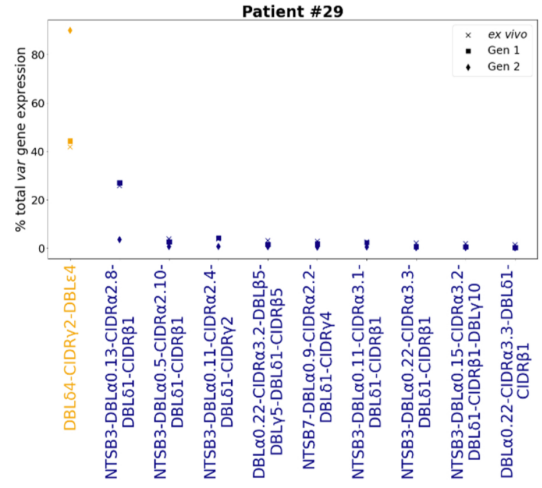

Patient #5
% total var gene expression
ex vivo
Gen 1
DBLα0.1-CIDRα3.1-DBLδ1-CIDRγ2
NTSB3-DBLα0.11-CIDRα2.4-DBLβ5-DBLδ1-CIDRβ1
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRβ5
NTSB3-DBLα0.5-CIDRα2.9-DBLδ1-CIDRγ2-DBLδ1
DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLz1-DBLε5
DBLα0.5-CIDRα2.3-DBLδ1-CIDRβ3
NTSB3-DBLα0.22-CIDRα3.1-DBLδ1-CIDRγ5
NTSB3-DBLα0.15-CIDRα2.3-DBLδ1-CIDRβ1
NTSB3-DBLα0.18-CIDRα5-DBLβ5-DBLγ3-DBLZ4
NTSB3-DBLα0.13-CIDRα2.9-DBLδ1-CIDRβ1

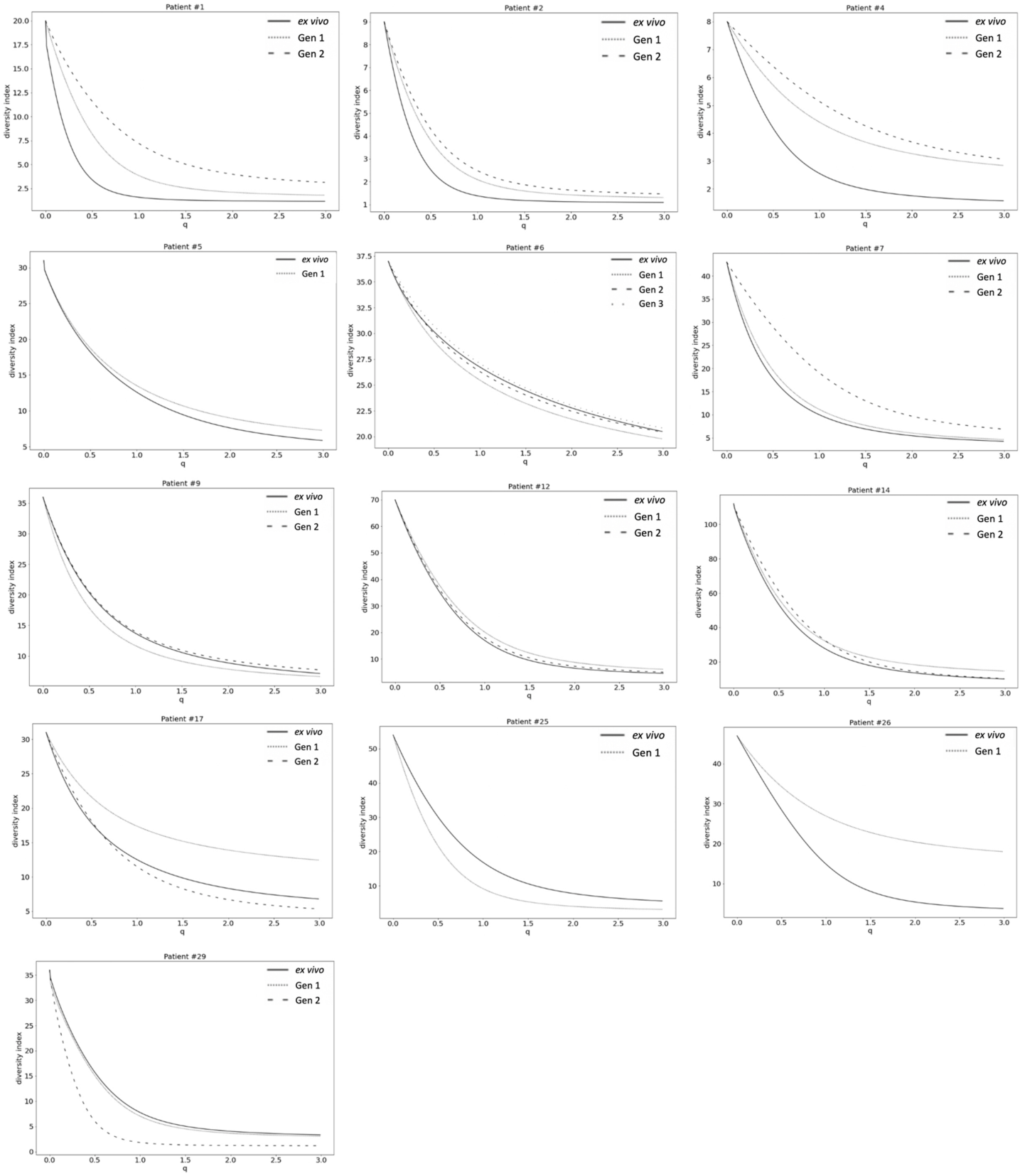

Patient #12
% total var gene expression
ex vivo
Gen 1
Gen 2
CIDRα3.1-DBLδ1-CIDRβ1
NTSB5-DBLα0.18-CIDRα3.1-DBLδ1-CIDRβ1
NTSB3-DBLα0.15-CIDRα3.2-DBLδ1-CIDRβ6
DBLα0.1-CIDRα3.1-DBLδ1-CIDRγ4
NTSB3-DBLα0.5-CIDRα2.5-DBLδ1-CIDRβ1
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRγ7
NTSB3-DBLα0.19-CIDRα2.8-DBLδ1-CIDRβ1-DBLγ10
CIDRα3.4-DBLδ1-CIDRβ1
DBLβ5-DBLγ13-DBLz2-DBLε4
NTSA1-DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8

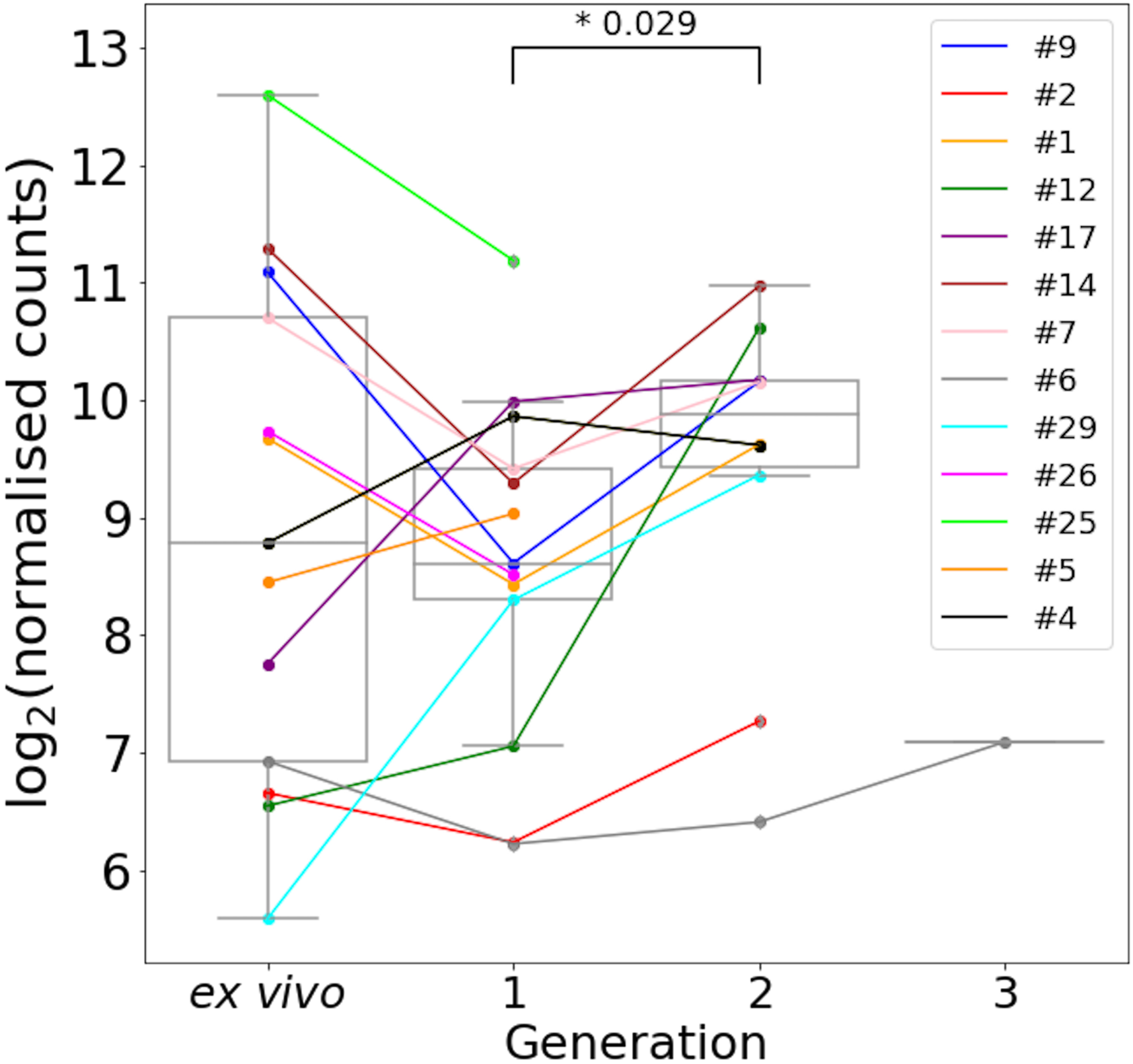

Patient #14
% total var gene expression
ex vivo
Gen 1
Gen 2
DBLα0.9-CIDRα2.7-DBLδ1-CIDRδ1
DBLβ5-DBLβ5-DBLδ1-CIDRβ1
DBLα1.5-CIDRβ4-DBLβ6-DBLβ6-DBLγ11-DBLγ4-DBLδ1
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRβ1-DBLγ10
DBLα0.17-DBLβ10-DBLβ5-DBLδ1-CIDRβ5
NTSB3-DBLα0.5-CIDRα2.3-DBLδ1-CIDRγ9
DBLα1.5-CIDRδ1-DBLβ3-DBLγ2-DBLγ4
DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLz2-DBLε5
CIDRδ1-DBLβ12-DBLγ2-DBLγ4-DBLδ1
NTSB3-DBLα0.6-CIDRα3.1-DBLβ5-DBLγ5-DBLδ1-CIDRβ5

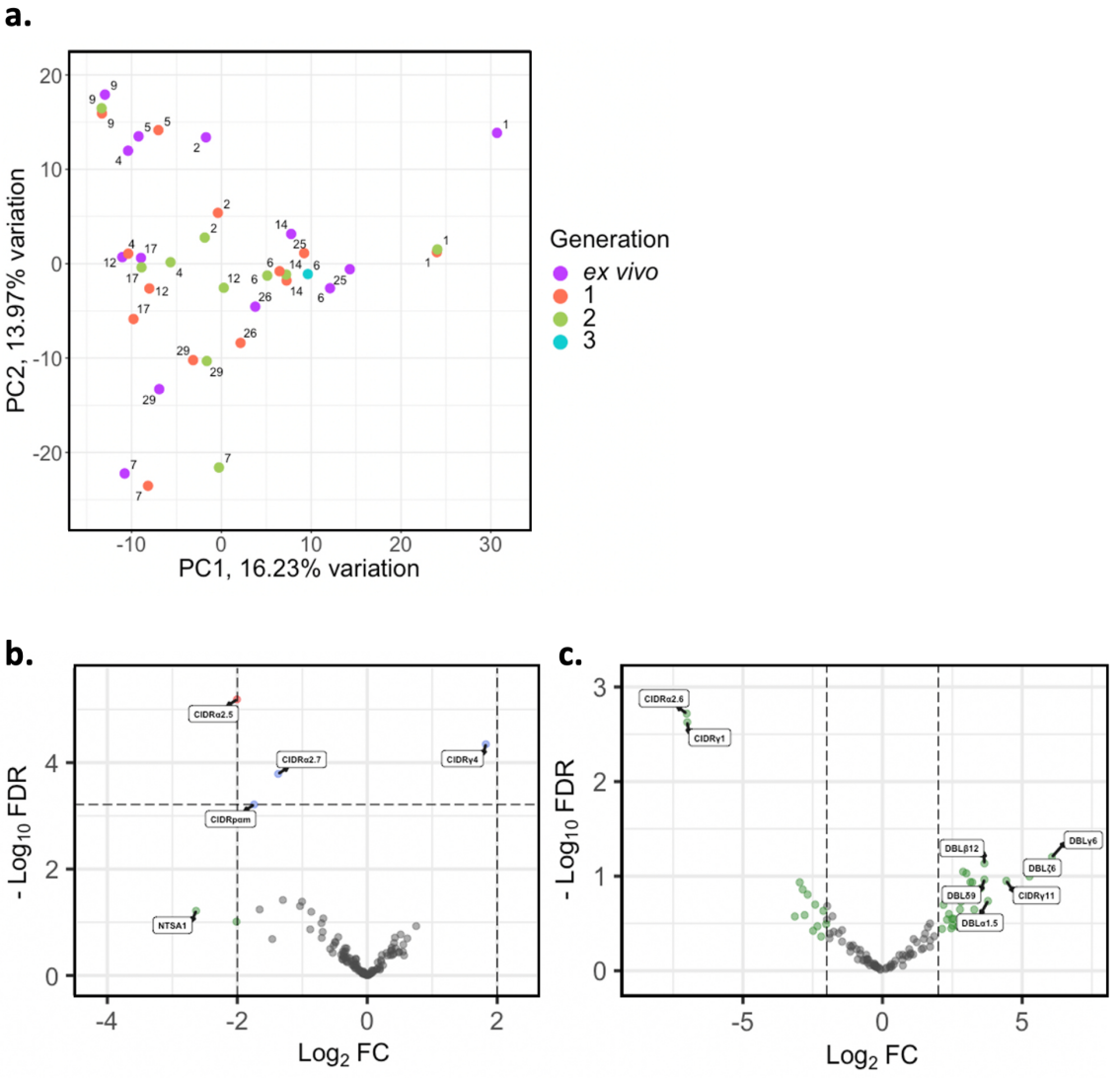

Patient #17
% total var gene expression
ex vivo
Gen 1
Gen 2
DBLγ2-DBLδ5-CIDRβ4-DBLz6-DBLε6
NTSB3-DBLα0.15-CIDRα3.2-DBLδ1-CIDRβ1
DBLγ11-DBLγ2-DBLδ1-CIDRγ12
NTSB3-DBLα0.16-CIDRα3.4-DBLδ1-CIDRβ1
DBLε1-DBLγ11-DBLδ1-CIDRβ6
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRβ1
NTSB3-DBLα0.5-CIDRα2.9-DBLδ1-CIDRβ6
NTSB3-DBLα0.9-CIDRα2.1-DBLδ1-CIDRβ1
NTSB3-DBLα0.9-CIDRα2.4-DBLδ1-CIDRγ7
NTSA1-DBLα1.1-CIDRα1.2-DBLβ11-DBLγ1-DBLε1-DBLγ8-DBLz1

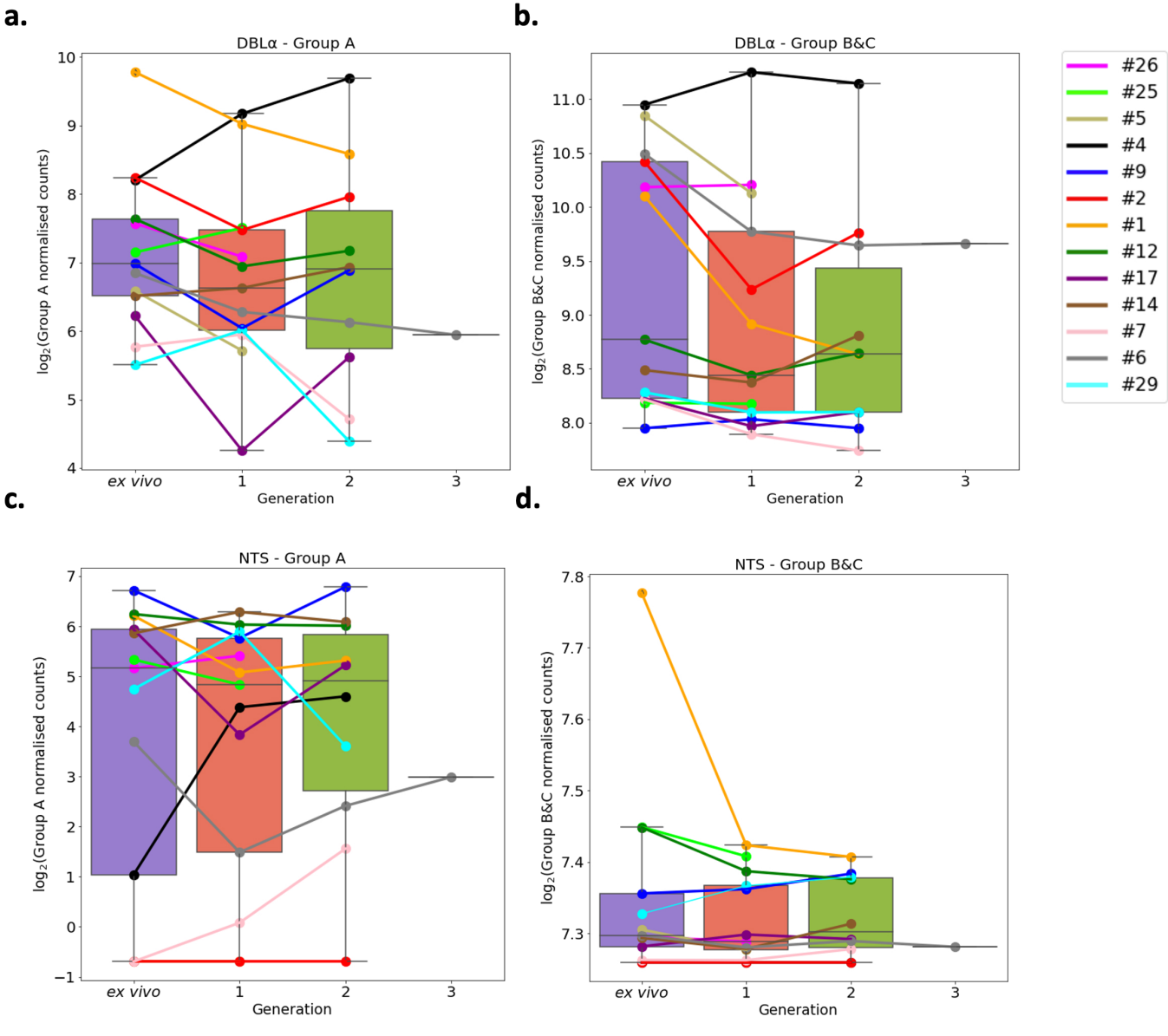

Patient #25
% total var gene expression
ex vivo
Gen 1
DBLβ5-DBLβ5-DBLβ5-DBLδ1-CIDRβ1
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRβ1
DBLα0.17-DBLβ5-DBLβ5-DBLδ1-CIDRβ1
NTSA5-DBLα1.6-CIDRδ1-DBLβ7-DBLβ3-DBLγ12-DBLγ11
DBLα1.5-CIDRβ4-DBLβ3-DBLγ2-DBLγ4-DBLδ1
NTSB3-DBLα0.18-CIDRα3.1-DBLβ5-DBLγ13-DBLδ1-CIDRγ9
NTSB3-DBLα0.15-CIDRα3.2-DBLδ1-CIDRβ1
DBLpam2-CIDRpam-DBLpam3-DBLepam4-DBLepam5-DBLε10
DBLα1.4-CIDRα1.7-DBLβ3-DBLγ2-DBLγ6-DBLδ5-DBLδ1-CIDRβ1
NTSB7-DBLα0.5-CIDRα2.6-DBLβ10-DBLδ1-CIDRβ1

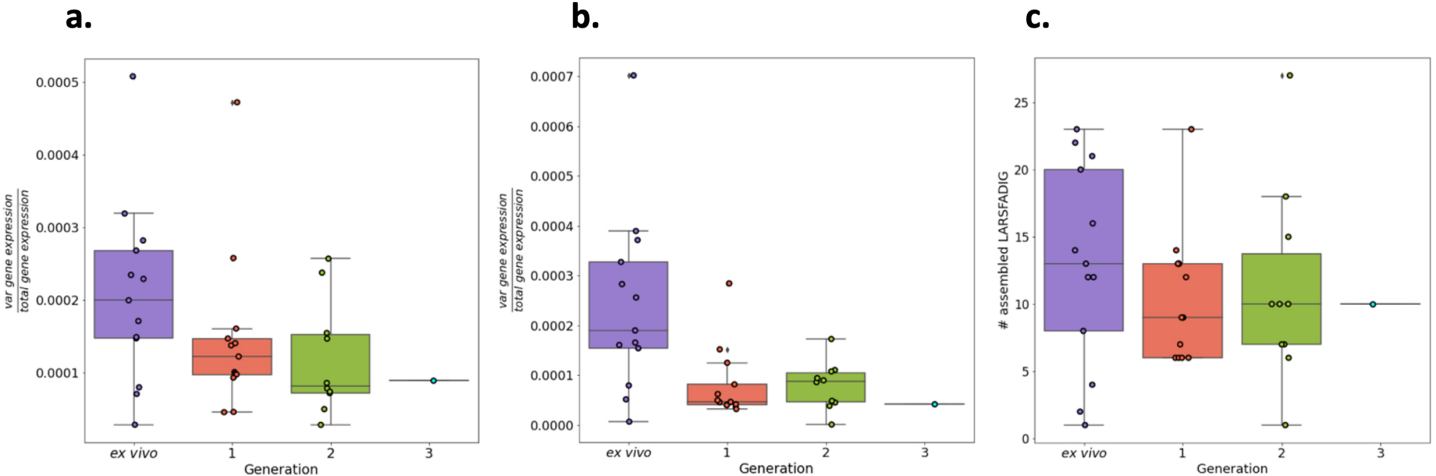

Patient #29
% total var gene expression
ex vivo
Gen 1
Gen 2
DBLδ4-CIDRγ2-DBLε4
NTSB3-DBLα0.13-CIDRα2.8-DBLδ1-CIDRβ1
NTSB3-DBLα0.5-CIDRα2.10-DBLδ1-CIDRβ1
NTSB3-DBLα0.11-CIDRα2.4-DBLδ1-CIDRγ2
DBLα0.22-CIDRα3.2-DBLβ5-DBLγ5-DBLδ1-CIDRβ5
NTSB7-DBLα0.9-CIDRα2.2-DBLδ1-CIDRγ4
NTSB3-DBLα0.11-CIDRα3.1-DBLδ1-CIDRβ1
NTSB3-DBLα0.22-CIDRα3.3-DBLδ1-CIDRβ1
NTSB3-DBLα0.15-CIDRα3.2-DBLδ1-CIDRβ1-DBLγ10
DBLα0.22-CIDRα3.3-DBLδ1-CIDRβ1

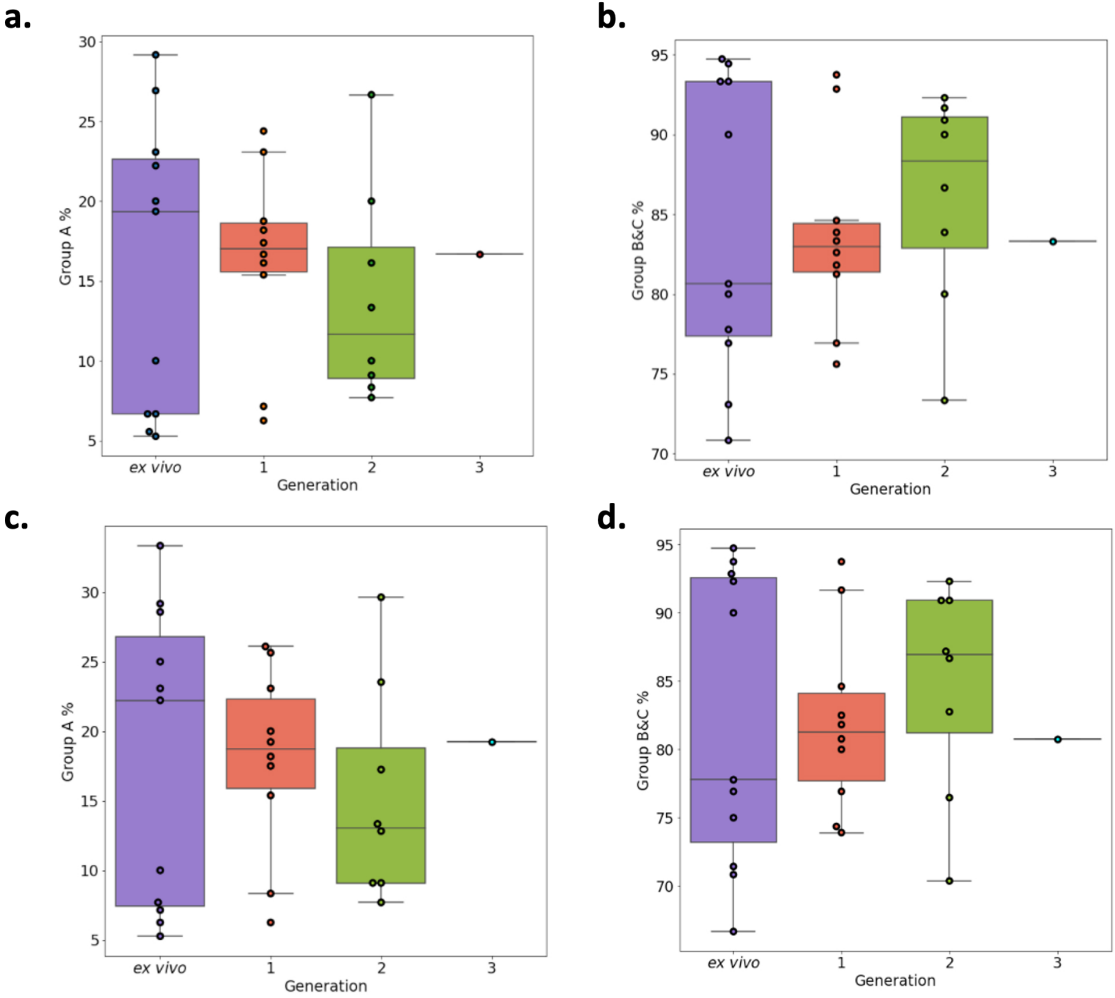

Patient #1
Patient #2
Patient #4
Patient #5
Patient #6
Patient #7
Patient #9
Patient #12
Patient #14
Patient #17
Patient #25
Patient #26
Patient #29
ex vivo
Gen 1
Gen 2
Gen 3
diversity index
q

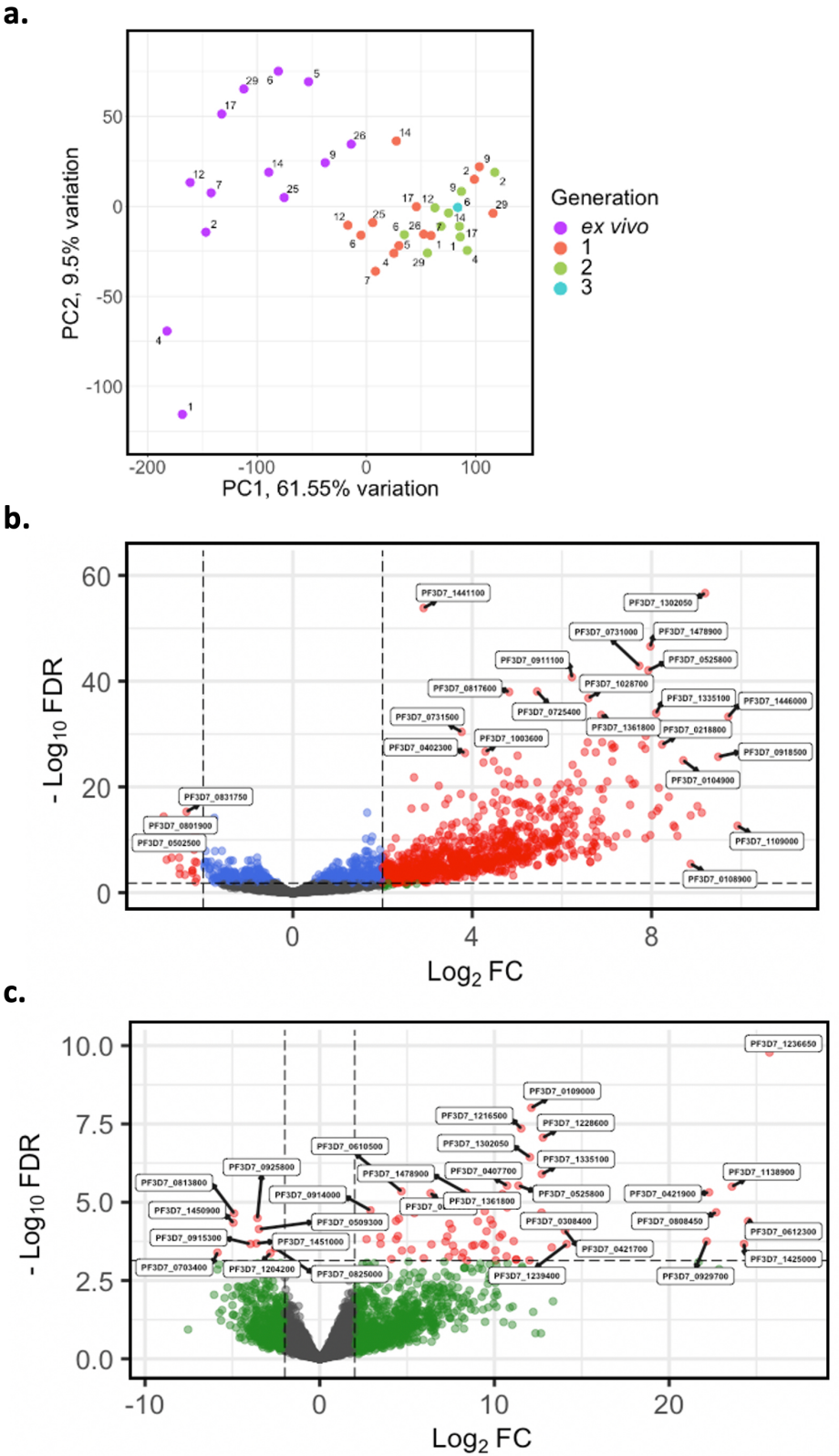

* 0.029
13
12
11
10
9
8
7
6
log2(normalised counts)
ex vivo
1
2
3
Generation
#9
#2
#1
#12
#17
#14
#7
#6
#29
#26
#25
#5
#4

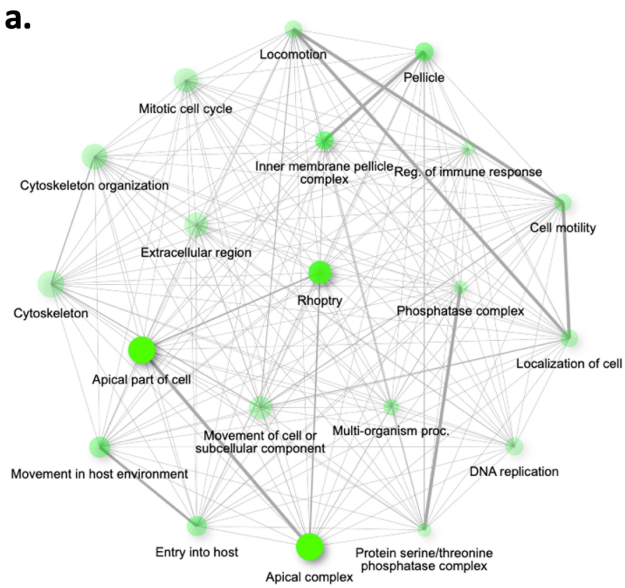

a.
PC2, 13.97% variation
PC1, 16.23% variation
Generation
ex vivo
1
2
3
b.
- Log10 FDR
Log2 FC
CIDRα2.5
CIDRα2.7
CIDRγ4
CIDRpam
NTSA1
c.
- Log10 FDR
Log2 FC
CIDRα2.6
CIDRγ1
DBLβ12
DBLγ6
DBLζ6
CIDRγ11
DBLδ9
DBLα1.5

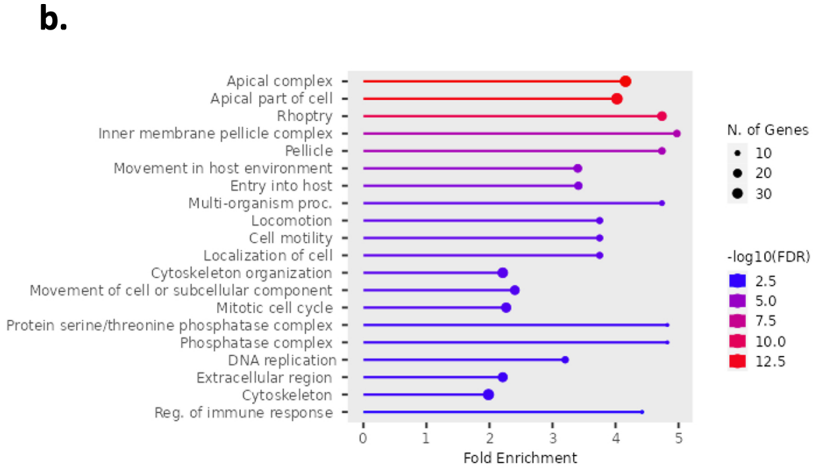

a.
DBLα - Group A
log2(Group A normalised counts)
ex vivo
1
2
3
Generation
b.
DBLα - Group B&C
log2(Group B&C normalised counts)
ex vivo
1
2
3
Generation
#26
#25
#5
#4
#9
#2
#1
#12
#17
#14
#7
#6
#29
c.
NTS - Group A
log2(Group A normalised counts)
ex vivo
1
2
3
Generation
d.
NTS - Group B&C
log2(Group B&C normalised counts)
ex vivo
1
2
3
Generation


a.
b.
c.
var gene expression / total gene expression
assembled LARSFADIG
Generation
ex vivo
1
2
3

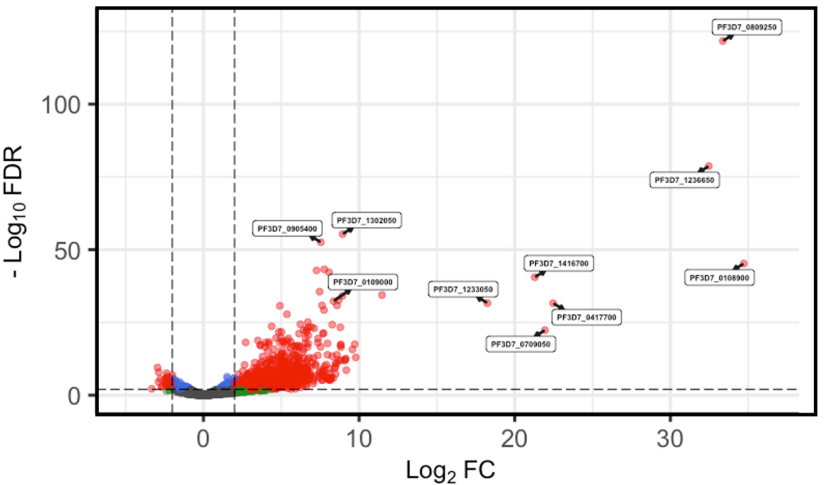

a.
Group A %
30
25
20
15
10
5
ex vivo
1
2
3
Generation
b.
Group B&C %
95
90
85
80
75
70
ex vivo
1
2
3
Generation
c.
Group A %
30
25
20
15
10
5
ex vivo
1
2
3
Generation
d.
Group B&C %
95
90
85
80
75
70
ex vivo
1
2
3
Generation


a.
PC2, 9.5% variation
PC1, 61.55% variation
Generation
ex vivo
1
2
3
b.
- Log10 FDR
Log2 FC
PF3D7_1441100
PF3D7_1302050
PF3D7_0731000
PF3D7_1478900
PF3D7_0911100
PF3D7_0525800
PF3D7_1028700
PF3D7_0817600
PF3D7_0725400
PF3D7_1335100
PF3D7_1446000
PF3D7_0731500
PF3D7_1361800
PF3D7_0218800
PF3D7_1003600
PF3D7_0402300
PF3D7_0918500
PF3D7_0104900
PF3D7_0831750
PF3D7_0801900
PF3D7_0502500
PF3D7_1109000
PF3D7_0108900
c.
- Log10 FDR
Log2 FC
PF3D7_1236650
PF3D7_0109000
PF3D7_1216500
PF3D7_1228600
PF3D7_0610500
PF3D7_1302050
PF3D7_1335100
PF3D7_0925800
PF3D7_1478900
PF3D7_0407700
PF3D7_1138900
PF3D7_0813800
PF3D7_0914000
PF3D7_1361800
PF3D7_0525800
PF3D7_0421900
PF3D7_1450900
PF3D7_0509300
PF3D7_0308400
PF3D7_0808450
PF3D7_0612300
PF3D7_0915300
PF3D7_1451000
PF3D7_0421700
PF3D7_1425000
PF3D7_0703400
PF3D7_1204200
PF3D7_0825000
PF3D7_1239400
PF3D7_0929700

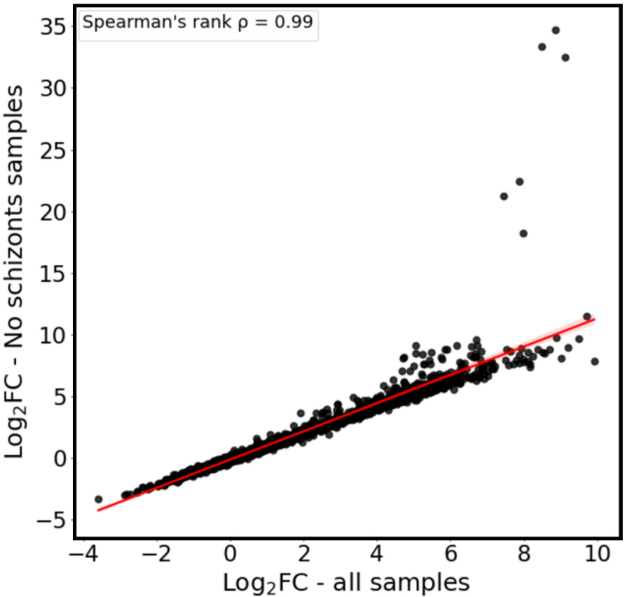

a.
Locomotion
Pellicle
Mitotic cell cycle
Inner membrane pellicle complex
Reg. of immune response
Cytoskeleton organization
Cell motility
Extracellular region
Rhoptry
Cytoskeleton
Phosphatase complex
Localization of cell
Apical part of cell
Movement of cell or subcellular component
Multi-organism proc.
Movement in host environment
DNA replication
Entry into host
Protein serine/threonine phosphatase complex
Apical complex

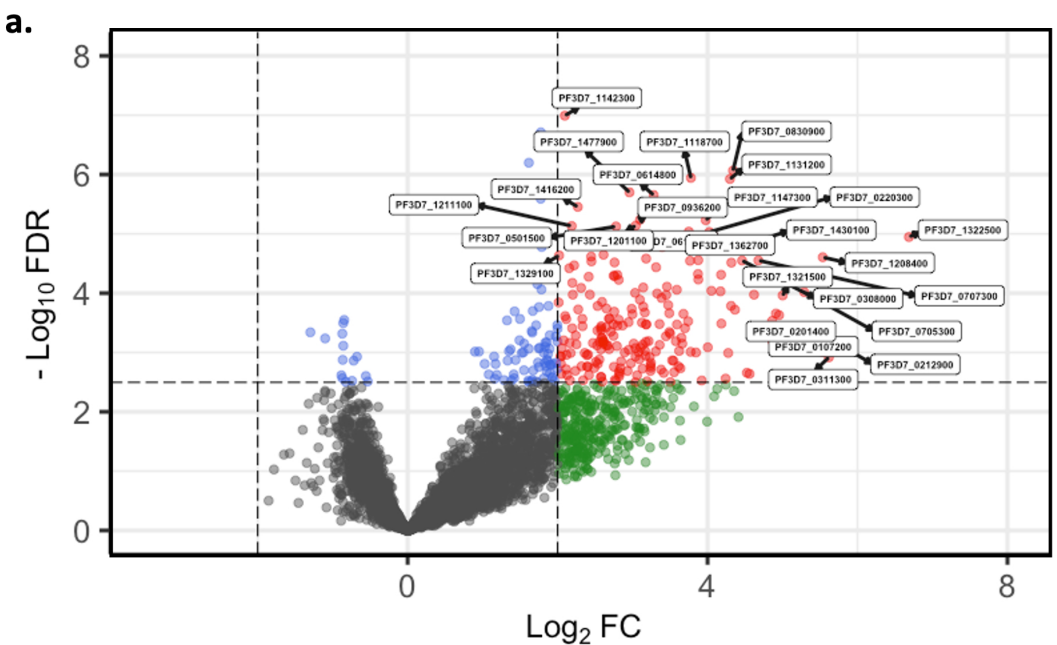

b.
Apical complex
Apical part of cell
Rhoptry
Inner membrane pellicle complex
Pellicle
Movement in host environment
Entry into host
Multi-organism proc.
Locomotion
Cell motility
Localization of cell
Cytoskeleton organization
Movement of cell or subcellular component
Mitotic cell cycle
Protein serine/threonine phosphatase complex
Phosphatase complex
DNA replication
Extracellular region
Cytoskeleton
Reg. of immune response
0
1
2
3
4
5
Fold Enrichment
N. of Genes
10
20
30
-log10(FDR)
2.5
5.0
7.5
10.0
12.5

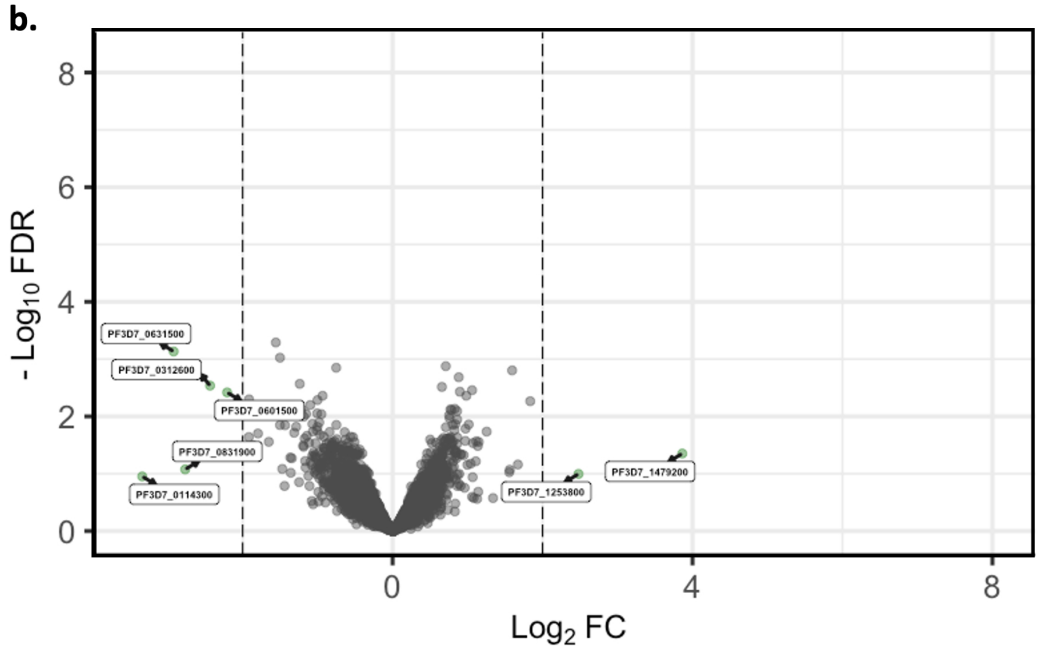
a.


PF3D7_0809250
PF3D7_1236650
PF3D7_0108900
PF3D7_1302050
PF3D7_0905400
PF3D7_0109000
PF3D7_1416700
PF3D7_1233050
PF3D7_0417700
PF3D7_0709050
100
50
0
- Log10 FDR
0
10
20
30
Log2 FC

b.


Spearman's rank ρ = 0.99
35
30
25
20
15
10
5
0
−5
Log2FC - No schizonts samples
−4
−2
0
2
4
6
8
10
Log2FC - all samples


a.
- Log10 FDR
Log2 FC
PF3D7_1142300
PF3D7_1477900
PF3D7_1118700
PF3D7_0830900
PF3D7_1131200
PF3D7_0614800
PF3D7_1416200
PF3D7_1147300
PF3D7_0220300
PF3D7_1211100
PF3D7_0936200
PF3D7_1430100
PF3D7_1322500
PF3D7_0501500
PF3D7_1201100
PF3D7_1362700
PF3D7_1208400
PF3D7_1329100
PF3D7_1321500
PF3D7_0308000
PF3D7_0707300
PF3D7_0201400
PF3D7_0705300
PF3D7_0107200
PF3D7_0212900
PF3D7_0311300


b.
- Log10 FDR
Log2 FC
PF3D7_0631500
PF3D7_0312600
PF3D7_0601500
PF3D7_0831900
PF3D7_0114300
PF3D7_1253800
PF3D7_1479200