# Sources of transcription variation in *Plasmodium falciparum*

Lindsey B. Turnbull[a], Katrina A. Button-Simons[b], Nestor Agbayani[b, c], Michael T. Ferdig[b, *]

[a] Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA

[b] Department of Biological Sciences, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, 46556, USA

[c] Rush School of Medicine, Chicago, IL, 60612, USA

[*] Corresponding author.

Email address: mferdig@nd.edu (M.T., Ferdig)

# Abstract

Variation in transcript abundance can contribute to both short-term environmental response and long-term evolutionary adaptation. Most studies are designed to assess differences in mean transcription levels and do not consider other potentially important and confounding sources of transcriptional variation. Detailed quantification of variation sources will improve our ability to detect and identify the mechanisms that contribute to genome-wide transcription changes that underpin adaptive responses. To quantify innate levels of expression variation, we measured mRNA levels for more than 5000 genes in the malaria parasite, *Plasmodium falciparum*, among clones derived from two parasite strains across biologically and experimentally replicated batches. Using a mixed effects model, we partitioned the total variation among four sources — between strain, within strain, environmental batch effects, and stochastic noise. We found 646 genes with significant variation attributable to at least one of these sources. These genes were categorized by their predominant variation source and further examined using gene ontology enrichment analysis to associate function with each source of variation. Genes with environmental batch effect and within strain transcript variation may contribute to phenotypic plasticity, while genes with between strain variation may contribute to adaptive responses and processes that lead to parasite strain-specific survival under varied conditions.

**Key words**: expression variation, malaria, transcription, cloning, microarray

# Introduction

Biology studies generally seek to identify and account for differences in phenotypic means between groups. In *Plasmodium falciparum*, analysis of mean expression, e.g. of each gene in a parasite sample to determine the differential mean expression for these genes between samples and across perturbations, has provided valuable insights into the regulation of gene expression across the lifecycle (Bozdech et al., 2003; Llinas et al., 2006), gene functions (Le Roch et al., 2003), transcriptional regulatory mechanisms (De Silva et al., 2008; Campbell et al., 2010; Painter et al., 2011), important clinical transcriptional phenotypes (Daily et al., 2007; Milner Jr et al., 2012) and the mechanism of action for antimalarial compounds (Mok et al., 2014; Siwo et al., 2015a). These studies have revealed a high adaptive capacity in the malaria parasite transcriptome in response to its environment. To more finely parse the sources of this adaptive capacity and to search for the mechanisms that underlie it, it is necessary to quantify the variation in transcripts beyond the differences in mean values. Focusing on sample means alone collapses all of the complex biological processes occurring among parasites within a culture or an infection to a single value, obscuring potentially important, biologically distinct information, including, the variation in transcription abundance among individual cells.

Studies in model organisms show the value of measuring the level of variation in expression across many genetically identical individuals that make up a population. For example, examination of *variation* in transcript abundance among single yeast cells within a population identified genes with expression changes based on cell sensing and adaptation to environmental events; these genes have a broad and heritable range of transcript abundances (Ansel et al., 2008). Differing environments can drive the abundance of specific transcripts, and can modulate the extent of transcript variation among individual organisms or between populations of organisms and can be measured as differences in transcript abundance of a specific gene(s) before and after perturbation (Acar et al., 2008; Keren et al., 2015). Organisms as diverse as *Arabidopsis* and *S. cerevisiae* have a subset of transcripts for which expression variation is genetically derived, differs among strains, and can be mapped to genes that control it (Ansel et al., 2008; Jimenez-Gomez et al., 2011).

For *P. falciparum*, whole genome transcription profiles have identified genes with variation among isogenic (clonal) parasites populations grown in identical environmental conditions (Scherf et al., 2008; Rovira-Graells et al., 2012; Reid et al., 2018). These three independent assessments revealed that approximately 5 percent of genes exhibit transcript variation among clones (genetically identical individual cells). Variation under these conditions was strongly correlated with binding to H3K9me3 and HP1, epigenetic marks that mediate reversible formation of heterochromatin to silence gene expression (Flueck et al., 2009; Lopez-Rubio et al., 2009; Gómez-Díaz et al., 2017) and to the potentially variable binding of transcription initiators (Reid et al., 2018). This research suggests that variation in transcript abundance, including variation in ApiAP2 transcription factors and their downstream transcripts (Martins et al., 2017), is regulated at the epigenetic level (Rovira-Graells et al., 2012). While these studies describe a previously uncharacterized regulatory mechanism of transcription variation, they did not partition the observed variation among potential sources. Specifically, they did not distinguish the contribution of strain variation from variation due to experimental batches and variation within isogenic clones.

Precise understanding of the basic biology of *P. falciparum* including, for example, the role of transcription in drug tolerance and resistance is clouded by the many possible sources that contribute to overall observed transcriptional variation for parasites at the same developmental stage and in the same physical condition. A more thorough understanding of this variation will improve our ability to distinguish signal from noise and perhaps illuminate new avenues for drug development and resistance prevention measures, e.g. antimalarial dosing regimens, and formulation of drug combinations. Partitioning the variation among relevant sources for each gene is an important first step. For example, knowing whether transcriptional variation in target genes is significantly impacted by the growth environment (e.g. pH differences or the presence of reactive oxygen species), or that the transcriptional variation is genotype specific (i.e. in the absence of mean differences) would, at a minimum, lead to better experimental designs. This knowledge could also potentially lead to more precise targeting, prioritization and administration of antimalarial drugs. Even in controlled *in vitro* experiments, knowledge of how much of the variation is due to differences between strains, environmental batch effects, individual parasite responses within an isogenic culture, and how much residual variation is due to other stochastic or unmeasured features will provide insights into the responses of *P. falciparum* to perturbation.

To quantitatively assess the different sources that contribute to the total amount of variation in gene expression, we used microarray-based gene expression data collected for multiple clones of two *P. falciparum* strains, HB3 and Dd2, across two experimental growth dates (Fig. 1). Using a multilevel mixed effects model, we removed confounding variation related to experimental differences in parasite stage and partitioned the remaining variation in expression for each of the 5540 transcripts in the *P. falciparum* genome into four variation sources: between strain variation (strain), within strain variation (clone), environmental batch effects, and stochastic variation. Our model quantitatively partitions the total variation in transcript abundance into these four sources. While disparate sources of variation are often broadly categorized as 'noise,' we use replicated batches of sub-cloned parasites from two different strains (Fig. 1) to show that there is important information about the underlying biology of *P. falciparum* hidden in the noise.

## Results

Malaria parasites are well known to have a highly regulated cascade of transcription across the erythrocytic lifecycle. The expression pattern of most genes approximates a sinusoidal curve with a distinct maximum and minimum level during the cyclical erythrocytic cycle (Bozdech et al., 2003; Llinas et al., 2006). Samples containing poorly synchronized parasites or representing a different time in the developmental lifecycle (at a different point in the sinusoidal curve) could increase the overall amount of observed variation and, if not accounted for, significantly impact the data interpretation. Consequently, to assess the success of our experimental design in controlling for both developmental stage and synchrony, we first compared our whole genome in vitro transcription profiles to time-course data from 3D7 taken across the erythrocytic lifecycle (Fig. 2). By correlating our data across the developmental time-course, it is clear that samples collected on 03/13 all are highly synchronized and were collected in the 16-20 hours post invasion (hpi) window. However, some samples collected on 07/14 correlated most closely with an earlier time point of 8-10 hpi. Thus, while rigorous standard synchronization methods were used to collect ‘ring stage’ parasites, a subset of 07/14 samples were significantly offset in their cell cycle window. To account for and mathematically remove transcription variation due to differences in parasite staging, a variable for stage specific variation was included in the model (Equation 2).

After accounting for stage-based variation, our model sequentially assessed genotype, batch, and clone-based transcriptional variation. The gene expression residual variation (Var($\varepsilon_{ij}$), Equation 2), which accounts for the variation around the mean among the samples, across all genes in the genome ranged from –3.03 to 5.54 with a median of zero. During the partitioning of variation among sources of interest, Var($\varepsilon_{ij}$) decreased based on the contribution of each source to the overall amount of variation. We identified 641 genes and 5 noncoding RNAs with statistically significant transcriptional variation due to one or more source. The remaining Var($\varepsilon_{ij}$) after accounting for genotype, environment, and clone-based variation ranged from –2.92 to 2.18 for these 646 genes with a median of zero. For the remainder of the genes in the genome ($n$ = 4894), which did not have significant variation based on one of the identified sources of interest, the range for Var($\varepsilon_{ij}$) was –3.49 to 5.54 with a median of zero.

## Model performance

Model fit statistics and adjusted p-values for each gene for each factor are reported in Table S1. To confirm the model performance, several genes were individually visualized and assessed as each source of variation was added to the model. For example, both PF3D7_0831700 (HSP70), and PF3D7_1415800 (a putative RNA methyl transferase) showed large overall variation in transcript abundance (Fig. 3). Our model discerned and effectively partitioned significant sources contributing to this variation at the level of genotype and environment. Fig 3A-3J shows the successive portioning of variation in transcript abundance through plotting residual variation as each source is added to the model. For PF3D7_0831700, the genotype-specific variation source was clearly distinguishable (A), whereas for PF3D7_1415800, no visible differences were observed for variation between HB3 and Dd2. Of note, the model was run on all samples simultaneously; however, we opted to display the genotypes (HB3 and Dd2) separately on the x-axis to emphasize this difference (see Methods for further model description). For both genes, removing stage-based variation did not substantially impact the variation among samples or between genotypes (Fig. 3B and 3G). When genotype was included in the model, and thus separated out from the other sources of variation (Fig. 3C and 3H), the mean level of residual transcript variation for PF3D7_0831700 was the same for strains HB3 and Dd2 (C) indicating that genotype was a significant source of overall variation in transcript abundance for this gene (*P* < 1.0E-13); this is typically credited to differential expression based on parasite isolate/strain. We observed no change to the residuals for PF3D7_1415800 (Fig3H), demonstrating that genotype did not contribute to the overall variation for this transcript. When environmental batch effect-based variation was added to the model (Fig. 3D and 3I), there was no significant change in the residuals for PF3D7_0831700 (D), indicating that for our study differences in the conditions between environmental batches did not contribute significantly to the variation for this particular gene. However, for PF3D7_1415800 (I) a significant reduction in residual variation was observed, indicating that environmental batch effects were a significant contributor to the overall variation for this gene (*P* = 3.72E-11). Finally, when clone was added as a source of variation to the model (Fig. 3E and 3J), the range of the remaining residual variation did change, but this was not statistically significantly for either depicted gene. This indicates that clonally-based sources of variation did not substantially contribute to the overall variation for these transcripts. Notably, for 12 genes not presented in Fig. 3, clone was the primary source of variation. In the full model, we also observe that the effect size of genotype-based variation for PF3D7_0831700 is 0.98 and for PF3D7_1415800 the effect size for l batch effect-based variation is 0.83, indicating that most of the variation among samples is due to a single source. While all transcript variation has been accounted for among the included sources for PF3D7_0831700,  only 86% of the total variation has been accounted in gene PF3D7_1415800; our model attributed this remaining transcriptional variation to stochastic noise.

## Most genes have one significant source of variation

While Fig. 3 closely examined two representative genes, our model comprehensively considered the sources of variation for all genes across the entire transcriptome. For each of the 5540 transcripts, total variation was partitioned (Fig. 4A). The effect size of these sources represented zero to one hundred percent of the total variation observed among the genes with a median effect size across all genes of 0.21 for stage, 0.002 for genotype, 0.42 for batch effects, and zero for clone (Fig. 4B). After controlling for stage, 641 genes and 5 noncoding RNAs exhibited statistically significant transcriptional variation attributable to one or more sources of variation included in our model. Among the 646 transcripts with significant variation due to any source, most had a single statistically significant source of transcriptional variation (Fig. 5; Table S2). The large majority of these genes derived transcriptional variation from environmental batch effects (472). A smaller number of genes varied based on genotype (171) and only a few genes (12) exhibited significant clone-based variation. Among the noncoding RNAs three had significant transcript variation based on experimental batch, one based on genotype, and one based on clone. Nine genes had statistically significant transcriptional variation attributable to more than one source; seven of these had variation by both genotype and environmental batch, and two had variation due to both genotype and clone.

To determine whether chromosomal location or structure contributed to transcriptional variation for any of the sources investigated, we mapped the genomic locations of transcriptionally variant genes onto their genomic locations. There were no significant patterns or enrichments for variable genes based on source, chromosome number, or chromosomal location (Fig. S1).

## Functional Enrichment of Genes by Source of Variation

To determine whether genes that shared a primary source of transcriptional variation shared biological functions, gene ontology (GO) enrichment analyses of molecular functions and biological processes were performed on each category of genes reaching significance for variance due to genotype, environment, and clone (Table. S3).

Nearly 30% of genes with significant genotype-based variation (53/171) belong to multi-gene families including: *rifin, stevor,* and *phist*. As such, the most significantly enriched GO terms included adhesion to host, regulation of erythrocyte aggregation and antigenic variation (GO: 0044406, 0034118, and 0020033). For these genes, involved in host evasion, each parasite only transcribes and express one transcript from the many potential genes in each family leading to expected transcription variation of individual transcripts within a population of parasites. While approximately half of the genes had higher mean transcriptional abundance HB3 and half in Dd2, the coefficient of variation (CoV, variance divided by the mean, and thus decoupled from differences in means), was at least 2-fold higher in HB3 than Dd2 for half of these host response genes (24/53) compared to twenty percent of genes (11/53) for which the CoV for Dd2 was higher (Table S4). Therefore, the observed level of genotype-based variation in host response genes could indicate differences in transcription regulatory mechanisms for host response genes between these two parasite strains.

Although only two genes had significant genotype-based *and* clone-based variation (PF3D7_0935400 and PF3D7_1302100), similar functions and processes were enriched in the sets of genes that had only genotype-based *or* clone-based variation. For example, our observation that adhesion, and aggregation were enriched functions in the clone-based gene list of genes (i.e., non-sequence based differences among genetically identical cells such as epigenetic marks) is consistent with an earlier report that most genes exhibiting variation among identical clones are members of these variable multi-gene families (Rovira-Graells et al., 2012). Many of these multi-gene families are positive for H9K3me3 and HP1 heterochromatin marks that silence expression and increase levels of variation (Flueck et al., 2009; Lopez-Rubio et al., 2009). Thus, in addition to confirming prior findings about the variation in this category of genes, our data show that the amount of variance observed in highly variable gene families differs between genotypes. Consequently, both the mean and variation in transcript abundance of host response genes in malaria parasites differs based on strain, suggesting that the underlying genetic control of each of these is heritable and under selection.

We did not observe an enrichment of stress response-related functions in the list of genes with environmental batch effect-based variation. This may reflect the tightly controlled nature of our experimental conditions in which rigorous protocols were designed to limit batch effects and did not introduce specific perturbations. This is notably distinct from other studies that intentionally test the effects of perturbations on transcript abundance and variation; consequently, our list of environmental batch-based variant genes is unlikely to contain genes with stress-related functions, reflecting the design of the experiment.

Threonine-type endopeptidase activity was the only enriched environmentally variant molecular function (GO: 0004298). Endopeptidases are responsible for breaking proteins apart at specific amino acids for recycling of these molecular building blocks. Even non-stressful differences in experimental environments may vary in the amounts of accessible resources, thus high innate levels of variation in genes such as protein cleavage enzymes could provide a way for parasite populations to tune in to their environment, stressful or not, in real-time, and this type of innate plasticity has been observed in other organisms (Gasch et al., 2000; Girardot et al., 2004). Most amino acid building blocks for malaria protein catabolism, including threonine, are derived by metabolizing host erythrocyte hemoglobin. Differences in the hemoglobin level of the host could promote differential expression of endopeptidases. While hemoglobin content within the cultures was not specifically measured, our data indicate differences in hemoglobin content may contribute to the observed environmental batch effect-based transcription variation of endopeptidase genes. For example, PF3D7_0931800, a subunit of the proteasome with threonine-type endopeptidase activity, had a mean abundance of 2317.9 and a CoV of 73.0 on our first sampling date. On the second sampling date, this transcript had a mean of 911.6 and a CoV of 31.7, exhibiting a strong shift in the mean, with the first experimental date having 2.54-fold higher transcript abundance (*P* = 2.44E-10, Bonferroni adjusted; Table S4).

## Transcriptional and Translational Expression Variation

Given their overall importance in regulating transcript abundance, we assessed the contribution of genotype to the transcriptional variation in genes involved in transcription and translational pathways. Differences in mean abundances or in the level of variation between genotypes in these transcripts could impact many downstream genes and result in substantial differences in gene expression regulation between the two parasite strains. We observed genes with genotype-based variation that are involved in transcription and translation. Variant genes involved in transcription including PF3D7_0522200 (*P* = 0.00016), a subunit of the general transcription factor, TFIID, had lower variation and lower mean expression in the drug resistant parasite Dd2. However, genes involved in translation, such as those that encode for ribosomal proteins PF3D7_0710900 ($P$ = 0.01878) and PF3D7_1223900 ($P$ = 0.02767), exhibited significantly higher variation and higher means in Dd2.

While not noted in the gene ontology enrichment processes, several genes involved in transcription and translation were significantly variant based on environmental batch. Variant genes that could influence transcription included the AP2 transcription factor domain gene PF3D7_1429200, and genes regulating chromatin condensation (PF3D7_0403100, PF3D7_0711500) and histones (PF3D7_1224500, PF3D7_1355300). Several genes involved in translation were also variant based on environmental batch including translation initiation factors (PF3D7_0907600, PF3D7_1250600, PF3D7_1332800), tRNA genes (PF3D7_1105700, PF3D7_1315700), and RNA polymerases (PF3D7_0708100, PF3D7_0205500, PF3D7_0303300, PF3D7_1134700, PF3D7_1213700).

## Discussion

This study dissects the biological sources that contribute to variation in gene expression in the malaria parasite. While most of the current scientific literature reports mean transcript abundance and dismisses variation as ‘noise,’ this study demonstrates that several biologically important sources contribute to the overall variation. By using a model to partition total variance, we assessed the quantitative contribution of each source for each transcript in the genome including stage, strain, environmental batch, clone, and stochastic variation. This partitioning provided a global and inclusive perspective of variation whereby the mechanisms underlying variation for each gene could be further investigated and better understood. Partitioning variation in this way required replication of samples for each source. This highly replicated data set included a total of 38 samples to parse variation across two stages within the lifecycle, two parasite strains, two different environmental batches, and among four subclones. Each of the identified sources of variation has important biological contributors and implications.

The transcript abundance of many genes throughout the *P. falciparum* genome is highly dependent on the number hours post- red blood cell invasion of the parasite during the erythrocytic cycle (Llinas et al., 2006; De Silva et al., 2008; Campbell et al., 2010). Differences in parasite developmental stage within the erythrocytic lifecycle can significantly confound results. In our study, clones were tightly synchronized such that more than 90% of parasites where within a 4h developmental window. RNA was collected 24h after thin smears consisted predominantly of highly segmented schizonts. While this study design should have produced samples that corresponded to the same developmental stage, correlations across samples and to a commonly referenced set of samples taken across the lifecycle revealed distinct stage differences in our samples. We controlled for these stage differences by adding an additional source of variation to our model. Variation in transcript abundance due to stage was accounted for and quantitatively removed prior to assessing the amount of variation due to any of the other sources.

After adjusting for stage variations, we observed that most variant genes have a single predominant source of variation, with the most prominent source being environmental batch. While previous reports suggested that 5% of the *P. falciparum* transcriptome is variant (Rovira-Graells et al., 2012; Reid et al., 2018), we found 646 transcripts, encompassing 12.0% of the transcriptome (including five non-coding RNAs) had significant expression variation *in vitro*. By including environmental batch as a source of variation in our study, something that to our knowledge has not been done, we detected more genes with transcriptional variation than previous reports. Environmental batch effect-based variation was observed in 472 genes. This variation can result from the same parasite (genotype, clone) having altered expression due to differences in the external conditions. These conditions will vary to some extent even in well-controlled experiments. For our study, potential variations in environmental batch conditions include red blood cell donor, media batch, parasitemia, and many other subtle and not directly controllable differences between experimental replicates/batches. Interestingly, we did not observe an enrichment of variant genes based on chromosomal location. Genes that were variant by genotype, environmental batch, and clone were dispersed throughout the genome and were not associated with subtelomeric regions or internal hypervariable regions. This suggests that transcriptional variation is not merely a product of currently understood structural genomic variability mechanisms.

Environmental batch effects are important to understand because they may not uniformly impact every gene. The impact of different environments or experimental conditions may manifest as large amounts of expression variation in some genes, while there will be no change in variation for other genes. Little is currently known about environmental sensing and adaptation to different 'normal' environments by malaria parasites. Genes with environmental batch-based variation may have an important role in the environmental tuning processes. Additionally, genes that have shifts in mean expression or variation in response to subtle environmental cues are commonly overlooked in standard expression studies and can contribute significantly to batch effects if not considered during experimental design. We identified several genes involved in transcriptional and translational processes with environmental batch-based variation. Variation in these genes could propagate further throughout the transcriptome by affecting specific downstream genes through transcription factor binding, or influencing transcription variation broadly through variations in the timing of translation initiation.

Excluding environmental batch-based variation from the study design, even under tightly controlled experimental conditions, can make the interpretation of results difficult as differences due to batch effects can be erroneously attributed to differences among genotypes or treatments. These results should encourage experimental procedures that are keenly aware of variation sources and study designs that allow for their accounting; these approaches will enhance the ability to detect differences due to the intended perturbation.

Strain-based variation can result from differences in the base pair sequence of a gene or gene regulator such as a copy number variations (Stranger et al., 2007; Eastman et al., 2011; Miles et al., 2016). Among the four variation sources assessed in our experiment, this is the most widely researched and easiest type of variation to identify. We measure strain-based variation more precisely by accounting for stage and microenvironment variation in the same model with comparisons of expression variation levels per gene among HB3 and Dd2 clones. Clone-based variation occurs among cloned individuals with identical genotypes and may reflect deviations in the way the transcription machinery interacts with epigenetic features of a gene. While the DNA sequence of a gene in two genetically identical cells is the same, other differences in epigenetic marks among individual cells that have been identified in *P. falciparum* could include base pair methylation, histone acetylation, nucleosome positioning, or other epigenetic marks (Ay et al., 2015). These epigenetic marks create variation by either altering the amount of time required to transcribe a sequence —resulting in more or fewer total copies, or by making the genetic sequence unavailable to transcription machinery thereby silencing gene expression. We have separated clonally-based variation from the total variation by taking whole transcriptome measurements of several first and second round clones with identical genotypes.

When only considering coding genes with significant variation by genotype or by clone, our study identified 3.2% (171 genes) and 0.2% (12 genes) of the transcriptome as variant. The most significantly enriched functions for genes with genotype and clone-based transcript variation involved the parasite's response to the human host. Thirty-two percent and fifty percent of genes variant by genotype and clone respectively belonged to multi-gene families including *var*, *rifin*, *stevor*, *and phist* gene families. These multi-gene families are well-known for their highly variable expression which aids in immune evasion (Gardner et al., 2002; Scherf et al., 2008). Aside from multi-gene families, one potentially interesting gene involved in transcription had significant strain-based variation. One of the subunits for TFIID (PF3D7_0522200) had substantial different transcriptional variation between the HB3 and Dd2 strains. While the absence of RNA Pol II-associated TFIID binding in Plasmodium makes the role of TFIID unknown and likely different in malaria than in other eukaryotes (Callebaut et al., 2005) this findings is worth investigating in the future.

The two parasite strains used in this study have substantially different drug resistant phenotypes. HB3 was originally isolated from Central America and has a high level of sensitivity to most antimalarial compounds. Dd2 was originally isolated from Southeast Asia after the emergence and fixation of chloroquine (CQ) resistance and underwent subsequent *in vitro* drug pressure with mefloquine which selected for a high level of resistance to 4-amino quinolone drugs (Oduola et al., 1988). The extended haplotype surrounding the causal mutation for CQ resistance, high level of linkage disequilibrium, and the lack of other haplotypes in Southeast Asia are hallmarks of a strong selective sweep (Wootton et al., 2002; Anderson, 2004; Su and Wootton, 2004). Based on the large overall impact CQ selection had on the genome and the transcriptome (Siwo et al., 2015b) of *P. falciparum*, and models based on laboratory experiments in *E. coli* that indicate strong selection increases variation (Ito et al., 2009; Eldar and Elowitz, 2010), we hypothesized that Dd2 would have higher levels of transcriptome-wide variation than HB3.

We found that, while genes with genotype-based variation have differences between means and genes with clone-based variation have differences in the CoV, these differences occurred on a gene-by-gene basis with neither HB3 nor Dd2 having overall higher levels of variation across the majority of genes for any source. However, when we investigated the differences between HB3 and Dd2 based on gene function, we found that variant genes involved in processes of antigenic variation and heat shock protein binding typically had larger transcriptional variation in HB3. This suggests that in non-stressful environments drug sensitive parasites have more variability in the amount of host response and stress response transcripts. As both antigenic variation and heat shock protein binding are important for response to the human immune system, genes with these functions are likely to be involved in the core environmental stress response for *P. falciparum* though further research is warranted to determine whether the amount of transcriptional variation in these genes changes during or after perturbations.

Our results suggest a different possibility for genes involved in transcription regulatory processes. These functional categories of genes associated with growth and proliferation are transcriptionally repressed during stress in yeast (Gasch et al., 2000; Causton et al., 2001). We identified several genes involved with transcription and translation with significant transcriptional variation between two different standard and non-stressful environments including transcription factor AP2-O3 (PF3D7_1429200). While differences in transcription variation transcription factor associated with mosquito stage ookinete sexual stage (Modrzynska et al., 2017) are interesting to observe during the asexual trophozoite stage, we were unable to determine the persistence of this variation into the sexual stages the parasite. Though we are unable to determine in our experiments which components of this standard environment contributed to the observed variation in transcription and translation related genes, both means and variance in these and other genes with environmental batch-based variation tended to be higher for the first culture date (03/13) than for the second (07/14).

Larger overall abundance together with highly variant amounts of transcriptional and translational machinery among genetically identical cells is consistent with a bet-hedging strategy in which a clonal population of cells with variant transcriptional phenotypes would be better suited to respond to wider range of future conditions (Seco-Hidalgo et al., 2015). Similarly, more overall ribosomes, and more variation in the amount of translational machinery could allow some cells to respond more rapidly to perturbation by adjusting the rate of protein synthesis. Understanding the baseline level and contributing sources of variation in these genes is valuable for developing a comprehensive biological interpretation of the changes seen after perturbation.

While we have accounted for some of the more obvious and relevant sources of variation, additional unmeasured sources will remain, and this will include some degree of stochastic noise. For our study these include the differences in the precise amount of bio-available transcriptional molecules each cell has at any moment, the position of each cell in the overall microenvironment, and many others that cannot, as of yet, be experimentally controlled for. This source of noise also includes the variation in our ability to measure the expression level itself such as the binding kinetics between probes on our array and our sample, detection limits, and differences that are not robust to standard normalization analyses.

Here we show that the total amount of variation in transcript abundance can be parsed using a mathematical model in *Plasmodium*. This model is generalizable to other data sets and could be used to explore the sources of variation under differing conditions in malaria and other organisms. In particular, we show that for a number of genes in *P. falciparum,* genotype contributes significantly to the total variation. This is an important feature of the parasite's biology that has implications for the development of new drugs and drug combinations as these could be more or less potent in some areas of the world based on parasite genotype. Clone-based variation also contributed significantly to a small number of genes. Both genotype and clone-based variation may be heritable and can be further explored to determine the regulatory mechanisms of variation within the parasite.

## Materials and Methods

### Parasite culture

*P. falciparum* cultures of HB3 and Dd2 and parasite clones were grown under standard conditions as described by Trager & Jensen (Trager and Jensen, 1976). Briefly, parasites were thawed into 5mL of RPMI 1640 (Invitrogen, Carlsbad, CA) supplemented with 25mM HEPES, 370μM hypoxanthine, 0.5% Albumax II (Invitrogen, Carlsbad, CA), 0.25% sodium bicarbonate (Mediatech, Inc., Manassas, Va) and 0.01 mg/mL gentamicin (Invitrogen, Carlsbad, CA), and 5% hematocrit O positive red blood cells. Cultures were maintained at 37°C under an atmosphere of 90% $N_2$, 5% $O_2$, and 5%$CO_2$.

### Study Design

*P. falciparum* parasite lines HB3 and Dd2 were thawed from minimally passaged stocks previously expanded and frozen in 2002 and 2004 respectively. Cultures were established and cloned by limiting dilution in 96-well plates. Cloning plates were screened weekly by light microscopy and positive wells were transferred to individual culture flasks. Clones were frozen once they reached 1% parasitemia in 0.5mL aliquots. A single clone from HB3 and Dd2 was randomly selected to undergo additional secondary cloning. Parental lines and first round clones were thawed, grown, synchronized, and collected for RNA during the first experimental timepoint labelled 03/13. Parental lines, first round, and second round clones were thawed, grown, synchronized, and collected for RNA during the second experimental time point labelled 07/14. Indicated first and second round clones (Fig. 1) were grown in triplicate during the second experimental date.

These standard culture conditions were consistent across both culturing dates to prevent, in as much as possible, the addition of experimental noise. Parasites used on different dates, to assess differences in microenvironment, were thawed from the same passage of frozen parasite stocks. The primary difference between culture dates to which environmental variation can be attributed is the red blood cell donor.

Cultures were synchronized three times across two lifecycles according to each parasite strain's cycle time (Reilly Ayala et al., 2010) using 5% sorbitol. The first synchronization occurred during the mid-ring stage. Parasites were allowed to reinvade, and the second synchronization occurred one lifecycle and 3 h after the first – 53 h for HB3 and 47.6h for Dd2 (cycle times are 50 h and 44.6 h, respectively)(Reilly et al., 2007). The third synchronization occurred 8 h after the second. Culture volume was increased to 20 mL during the synchronization cycles, and parasites were monitored by thin smear microscopy for re-invasion every 2 h after the third synchronization. As highly segmented schizonts are morphologically distinct, and this stage lasts for less than 4 h, only cultures with more than 90% of parasites in the highly segmented schizont were used, and this point was designated time zero (T0). RNA was collected 24 h after T0.

The NF-54 parasite isolate, from which 3D7 was cloned, was also cultured and synchronized. Eight collections of this parasite occurred across a single asexual lifecycle at 2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46 hours. Cultures were washed with warm PBS, pelleted and flash frozen using liquid $N_2$ and stored at −80°C for less than 2 weeks prior to RNA extraction.

### RNA extraction and cDNA synthesis

Total parasite RNA was extracted from frozen red blood cell pellets using TriZol reagent (Invitrogen, Carlsbad, CA) as previously described (Bozdech et al., 2003). Quantity and quality of RNA was determined by Nanodrop (Nanodrop Technologies) and stored at -80°C. A total of 300 ng RNA was used as starting material for cDNA synthesis using the Sigma WTA2 whole transcriptome amplification kit (Sigma Aldrich, St. Louis, MO).

### cDNA labelling and hybridization to exon microarrays

A total of 1.0 µg cDNA was labeled using Cy3 dye attached to 65% A/T rich random hexamers (TriLink) as primers for cDNA synthesis by Klenow fragments (New England Biolabs). For the NF-54 reference, equal amounts of cDNA from each collection time were pooled prior to labelling. Then 2.5 µg of labeled cDNA was suspended in Agilent Expression Hybe Buffer and Blocking Agent (Agilent) and loaded onto whole transcriptome Agilent exon arrays (Turnbull et al., 2017). Hybridizations were incubated for 17h at 65°C at 12 rpm and washed according to standard protocols (Agilent). Multi-image TIFFs of the microarrays were obtained using a 2 µM scanner (Roche NimbleGen Inc., Madison, WI) and extracted using Agilent Feature Extraction software (Agilent).

## Data processing and normalization

The Agilent exon array consists of 62,976 probes which provide transcriptional abundance information for 5540 genes in the malaria genome and 100 noncoding RNAs (Turnbull et al., 2017). Probes with intensities less than 1.5 standard deviations of background were first trimmed from the probe sets. A random sampling for 1000 probe sets was used to determine the 5% false discovery rate (FDR) for each array. Probes below the 5% FDR cut-off were also excluded from further analysis. All samples were then quantile normalized together in robust multi-array averaging to adjust and standardize distributions for the study. Probe intensity values were then averaged and transcriptional abundances were reported by exon and by gene.

The transcriptional patterns of *P. falciparum* are highly associated with the progression through the asexual lifecycle from ring to schizont in which transcription of genes is turned on and off in a highly organized cascade. To account for known differences in asexual cycle time between the parasite strains used in this study and determine a normalized RNA expression level a pool of NF54 RNA from across the parasite lifecycle was run on the exon array and used as the denominator in a $\log_2$ ratio normalization. To further account for potential differences in the staging process of individual cultures, Spearman correlations of each $\log_2$ normalized whole genome transcription profile were then correlated to community standard profiles of 3D7 taken at hour intervals across the entire life cycle (Bozdech et al., 2003). The highest correlation value for each sample was used to determine the corrected hpi. Because subtle differences in stage during the trophozoite phase of the lifecycle can impact RNA abundance, and stage differences were not the focus of this investigation, this corrected hpi value was added to the model as an additional source of variation. $\text{Log}_2$ normalized values were used for assessing variation and stage-based variation was accounted for *prior to* mathematically partitioning the remaining variation among the other identified biological sources.

## Model for partitioning variation

Beginning from the base model, which includes the population mean (of clones measured) and the residual error (since we have not yet accounted for any sources of variation, the starting residual error includes all of the variation around the mean),

$$Gene\_Exp_i = \beta_0 + \varepsilon_i \qquad \text{(Equation 1)}$$

a random intercepts model was used to partition the starting residual error ($\varepsilon_i$) into two categories: variation due to stage ($s_i$), and residual unexplained variation due to all other sources ($\varepsilon_{ij}$).

$$Gene\_Exp_i = \beta_0 + s_i + \varepsilon_{ij} \qquad \text{(Equation 2)}$$

Here $\beta_0 + s_i$ represents the model's estimate of the mean expression level for stage *i*.

$\varepsilon_{ij}$ then retains the remaining variation not explained by stage. The total variation in gene expression can be represented by

$$Var\left(s_i + \varepsilon_{ij}\right) = Var(s_i) + Var(\varepsilon_{ij}) \qquad \text{(Equation 3)}$$

$Var(\varepsilon_{ij})$ can be further decomposed to account for expression differences due to parasite genotype.

$$Gene\_Exp_{ijk} = \beta_0 + s_i + g_{ij} + \varepsilon_{ijk} \qquad \text{(Equation 4)}$$

$\beta_0 + s_i + g_{ij}$ represents the models estimate of stage *i*'s mean gene expression value for genotype *j*. The total variation in gene expression can be represented by

$$Var(s_i + g_{ij} + \varepsilon_{ijk}) = Var(s_i) + Var(g_{ij}) + Var(\varepsilon_{ijk}). \quad \text{(Equation 5)}$$

$Var(g_{ij})$ provides an estimate of genotypic variation when parasites are tightly synchronized, and stage differences are controlled for.

The remaining variation now consists of variation due to environmental batch-based conditions, individual sub-clones, and unexplained variation. Next, we account for the contribution of environmental batch ($d_{ijk}$) to the unexplained variation in the residuals ($\varepsilon_{ijkl}$).

$$Gene\_Exp_{ijkl} = \beta_0 + s_i + g_{ij} + d_{ijk} + \varepsilon_{ijkl} \quad \text{(Equation 6)}$$

In Equation 6, $\beta_0 + s_i + g_{ij} + d_{ijk}$ represents the models estimate for the mean gene expression value for genotype *j* during stage *i* on a given date *k*. To this point, the model estimate variation in gene expression is thus represented by

$$Var(s_i + g_{ij} + \varepsilon_{ijk} + d_{ijk}) = Var(s_i) + Var(g_{ij}) + Var(d_{ijk}) + Var(\varepsilon_{ijkl}).$$

(Equation 7)

Environmental batch-based variation due to culture growth date has now been isolated as an important source of within strain variation. The only biologically identified source of variation left to account for in this study is based on individual sub-clones ($c_{ijkl}$).

$$Gene\_Exp_{ijkl} = \beta_0 + s_i + g_{ij} + d_{ijk} + c_{ijkl} + \varepsilon_{ijklm} \quad \text{(Equation 8)}$$

In our final model Equation 8, $\beta_0 + s_i + g_{ij} + d_{ijk} + c_{ijkl}$ gives the estimated amount of gene expression variation for a given clone *l* of a specific genotype *j*, on a given date *k*, during lifecycle stage *i*. And the total variation in gene expression becomes

$$Var\left(s_i + g_{ij} + \varepsilon_{ijk} + d_{ijk} + \ c_{ijkl}\right) = Var(s_i) + Var\left(g_{ij}\right) + Var\left(d_{ijk}\right) + \ Var(c_{ijkl}) + Var(\varepsilon_{ijklm}). \qquad \text{(Equation 9)}$$

$Var\left(c_{ijkl}\right)$ gives an estimate of variation in gene expression among clones that is not due to growth conditions, or genetic differences between genotypes HB3 and Dd2. Thus, the starting unexplained residual variation has been partitioned based on three biologically important sources: genotype, environmental batch (growth date), and clone. The remaining residual variation $Var(\varepsilon_{ijklm})$ gives the model's estimate for any unaccounted-for variation (*i.e.* noise).

For each source we determined if the partitioned amount of variation was significantly different from zero by calculating the observed difference in $-2\ln(likelihood)$ for the model without the random effect (source) with the model containing the random effect. We compared this value to a reference distribution $(\mathcal{X}_0^2 + \mathcal{X}_1^2)/2$ of expected likelihood differences to determine a $p$ value (Hruschka et al., 2005). The Bonferroni method was applied to account for multiple testing ($\alpha = 0.05$).

## Acknowledgements

LBT, KAB were supported by Eck Institute for Global Health Fellowships.

This work was supported by the Indiana Clinical and Translational Sciences Institute [UL1TR001108] from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award and with the support from the Center for Quantitative Biology [P50 GM071508]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Conflicts of Interest

The authors declare that there are no conflicts of interest with this work.

## CRediT authorship contribution statement

**Lindsey B. Turnbull**: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original draft, Writing - Review & Editing, Visualization, Funding acquisition. **Katrina A. Button-Simons**: Methodology, Software, Formal analysis, Data curation, Writing- Review & Editing. **Nestor Agbayani**: Investigation, Writing - Review & Editing. **Michael T. Ferdig**: Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding acquisition.

## References

Acar, M., Mettetal, J.T.,van Oudenaarden, A., 2008. Stochastic switching as a survival strategy in fluctuating environments. Nat Genet 40, 471-475.

Anderson, T.J., 2004. Mapping drug resistance genes in plasmodium falciparum by genomewide association. Current Drug Targets-Infectious Disorders 4, 65-78.

Ansel, J., Bottin, H., Rodriguez-Beltran, C., Damon, C., Nagarajan, M., Fehrmann, S., Francois, J.,Yvert, G., 2008. Cell-to-cell stochastic variation in gene expression is a complex genetic trait. PLoS Genet 4, e1000049.

Ay, F., Bunnik, E.M., Varoquaux, N., Vert, J.P., Noble, W.S.,Le Roch, K.G., 2015. Multiple dimensions of epigenetic gene regulation in the malaria parasite plasmodium falciparum: Gene regulation via histone modifications, nucleosome positioning and nuclear architecture in p. Falciparum. BioEssays 37, 182-194.

Bozdech, Z., Llinás, M., Pulliam, B.L., Wong, E.D., Zhu, J.,DeRisi, J.L., 2003. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. PLoS biology 1, e5.

Callebaut, I., Prat, K., Meurice, E., Mornon, J.-P.,Tomavo, S., 2005. Prediction of the general transcription factors associated with rna polymerase ii in plasmodium falciparum: Conserved features and differences relative to other eukaryotes. BMC genomics 6, 1-20.

Campbell, T.L., De Silva, E.K., Olszewski, K.L., Elemento, O.,Llinas, M., 2010. Identification and genome-wide prediction of DNA binding specificities for the apiap2 family of regulators from the malaria parasite. PLoS Pathog 6, e1001165.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S.,Young, R.A., 2001. Remodeling of yeast genome expression in response to environmental changes. Molecular biology of the cell 12, 323-337.

Daily, J.á., Scanfeld, D., Pochet, N., Le Roch, K., Plouffe, D., Kamal, M., Sarr, O., Mboup, S., Ndir, O.,Wypij, D., 2007. Distinct physiological states of plasmodium falciparum in malaria-infected patients. Nature 450, 1091.

De Silva, E.K., Gehrke, A.R., Olszewski, K., León, I., Chahal, J.S., Bulyk, M.L.,Llinás, M., 2008. Specific DNA-binding by apicomplexan ap2 transcription factors. Proceedings of the National Academy of Sciences 105, 8393-8398.

Eastman, R.T., Dharia, N.V., Winzeler, E.A.,Fidock, D.A., 2011. Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured plasmodium falciparum parasites. Antimicrob Agents Chemother 55, 3908-3916.

Eldar, A.,Elowitz, M.B., 2010. Functional roles for noise in genetic circuits. Nature 467, 167-173.

Flueck, C., Bartfai, R., Volz, J., Niederwieser, I., Salcedo-Amaya, A.M., Alako, B.T., Ehlgen, F., Ralph, S.A., Cowman, A.F., Bozdech, Z.*, et al.*, 2009. Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS Pathog 5, e1000569.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E.,Bowman, S., 2002. Genome sequence of the human malaria parasite plasmodium falciparum. Nature 419, 498.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D.,Brown, P.O., 2000. Genomic expression programs in the response of yeast cells to environmental changes. Molecular biology of the cell 11, 4241-4257.

Girardot, F., Monnier, V.,Tricoire, H., 2004. Genome wide analysis of common and specific stress responses in adult drosophila melanogaster. BMC genomics 5, 74.

Gómez-Díaz, E., Yerbanga, R.S., Lefèvre, T., Cohuet, A., Rowley, M.J., Ouedraogo, J.B.,Corces, V.G., 2017. Epigenetic regulation of plasmodium falciparum clonally variant gene expression during development in anopheles gambiae. Scientific reports 7, 40655.

Hruschka, D.J., Kohrt, B.A.,Worthman, C.M., 2005. Estimating between-and within-individual variation in cortisol levels using multilevel models. Psychoneuroendocrinology 30, 698-714.

Ito, Y., Toyota, H., Kaneko, K.,Yomo, T., 2009. How selection affects phenotypic fluctuation. Molecular systems biology 5, 264.

Jimenez-Gomez, J.M., Corwin, J.A., Joseph, B., Maloof, J.N.,Kliebenstein, D.J., 2011. Genomic analysis of qtls and genes altering natural variation in stochastic noise. PLoS Genet 7, e1002295.

Keren, L., Van Dijk, D., Weingarten-Gabbay, S., Davidi, D., Jona, G., Weinberger, A., Milo, R.,Segal, E., 2015. Noise in gene expression is coupled to growth rate. Genome research, gr. 191635.191115.

Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J*., et al.*, 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301, 1503-1508.

Llinas, M., Bozdech, Z., Wong, E.D., Adai, A.T.,DeRisi, J.L., 2006. Comparative whole genome transcriptome analysis of three plasmodium falciparum strains. Nucleic Acids Res 34, 1166-1173.

Lopez-Rubio, J.J., Mancio-Silva, L.,Scherf, A., 2009. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host Microbe 5, 179-190.

Martins, R.M., Macpherson, C.R., Claes, A., Scheidig-Benatar, C., Sakamoto, H., Yam, X.Y., Preiser, P., Goel, S., Wahlgren, M.,Sismeiro, O., 2017. An apiap2 member regulates expression of clonally variant genes of the human malaria parasite plasmodium falciparum. Scientific reports 7, 14042.

Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., Drury, E.,O'Brien, J., 2016. Indels, structural variation, and recombination drive genomic diversity in plasmodium falciparum. Genome research.

Milner Jr, D.A., Pochet, N., Krupka, M., Williams, C., Seydel, K., Taylor, T.E., Van de Peer, Y., Regev, A., Wirth, D.,Daily, J.P., 2012. Transcriptional profiling of plasmodium falciparum parasites from patients with severe malaria identifies distinct low vs. High parasitemic clusters. PloS one 7, e40739.

Modrzynska, K., Pfander, C., Chappell, L., Yu, L., Suarez, C., Dundas, K., Gomes, A.R., Goulding, D., Rayner, J.C.,Choudhary, J., 2017. A knockout screen of apiap2 genes reveals networks of interacting transcriptional regulators controlling the plasmodium life cycle. Cell host & microbe 21, 11-22.

Mok, S., Ashley, E.A., Ferreira, P.E., Zhu, L., Lin, Z., Yeo, T., Chotivanich, K., Imwong, M., Pukrittayakamee, S.,Dhorda, M., 2014. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science, 1260403.

Oduola, A.M., Milhous, W., Weatherly, N., Bowdre, J.,Desjardins, R., 1988. Plasmodium falciparum: Induction of resistance to mefloquine in cloned strains by continuous drug exposure in vitro. Experimental parasitology 67, 354-360.

Painter, H.J., Campbell, T.L.,Llinás, M., 2011. The apicomplexan ap2 family: Integral factors regulating plasmodium development. Molecular and biochemical parasitology 176, 1-7.

Reid, A.J., Talman, A.M., Bennett, H.M., Gomes, A.R., Sanders, M.J., Illingworth, C.J., Billker, O., Berriman, M.,Lawniczak, M.K., 2018. Single-cell rna-seq reveals hidden transcriptional variation in malaria parasites. Elife 7, e33105.

Reilly Ayala, H.B., Wacker, M.A., Siwo, G.,Ferdig, M.T., 2010. Quantitative trait loci mapping reveals candidate pathways regulating cell cycle duration in plasmodium falciparum. BMC Genomics 11, 577.

Reilly, H.B., Wang, H., Steuter, J.A., Marx, A.M.,Ferdig, M.T., 2007. Quantitative dissection of clone-specific growth rates in cultured malaria parasites. Int J Parasitol 37, 1599-1607.

Rovira-Graells, N., Gupta, A.P., Planet, E., Crowley, V.M., Mok, S., Ribas de Pouplana, L., Preiser, P.R., Bozdech, Z.,Cortes, A., 2012. Transcriptional variation in the malaria parasite plasmodium falciparum. Genome Res 22, 925-938.

Scherf, A., Lopez-Rubio, J.J.,Riviere, L., 2008. Antigenic variation in plasmodium falciparum. Annu Rev Microbiol 62, 445-470.

Seco-Hidalgo, V., Osuna, A.,De Pablos, L.M., 2015. To bet or not to bet: Deciphering cell to cell variation in protozoan infections. Trends in parasitology 31, 350-356.

Siwo, G.H., Smith, R.S., Tan, A., Button-Simons, K.A., Checkley, L.A.,Ferdig, M.T., 2015a. An integrative analysis of small molecule transcriptional responses in the human malaria parasite plasmodium falciparum. BMC Genomics 16, 1030.

Siwo, G.H., Tan, A., Button-Simons, K.A., Samarakoon, U., Checkley, L.A., Pinapati, R.S.,Ferdig, M.T., 2015b. Predicting functional and regulatory divergence of a drug resistance transporter gene in the human malaria parasite. BMC Genomics 16, 115.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., De Grassi, A.,Lee, C., 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, 848-853.

Su, X.Z.,Wootton, J.C., 2004. Genetic mapping in the human malaria parasite plasmodium falciparum. Mol Microbiol 53, 1573-1582.

Trager, W.,Jensen, J.B., 1976. Human malaria parasites in continuous culture. Science 193, 673-675.

Turnbull, L.B., Siwo, G.H., Button-Simons, K.A., Tan, A., Checkley, L.A., Painter, H.J., Llinás, M.,Ferdig, M.T., 2017. Simultaneous genome-wide gene expression and transcript isoform profiling in the human malaria parasite. PloS one 12, e0187595.

Wootton, J.C., Feng, X., Ferdig, M.T., Cooper, R.A., Mu, J., Baruch, D.I., Magill, A.J.,Su, X.-z., 2002. Genetic diversity and chloroquine selective sweeps in plasmodium falciparum. Nature 418, 320.

## Figure legends

**Fig 1.** Experimental Design for Partitioning Variation. Laboratory parasite strains HB3 and Dd2 were cloned to obtain four 1st round clones. One of these 1st round clones was further sub-cloned to obtain four 2nd round clones. HB3, Dd2 and 1st round clones were cultured twice to produce date replicated environmental batches. All 2nd round clones and biological replicates (*, *n* = 3) were cultured once during the second environmental batch. Different sources of variation can be assessed with the study design including environment batch, genotype, and clone-based variation as indicated by brackets. The final data set included 38 parasite lines as follows: a total of 20 HB3 cultures; a total of 18 Dd2 cultures; 10 cultures from 03/13; and 28 cultures from 07/14.

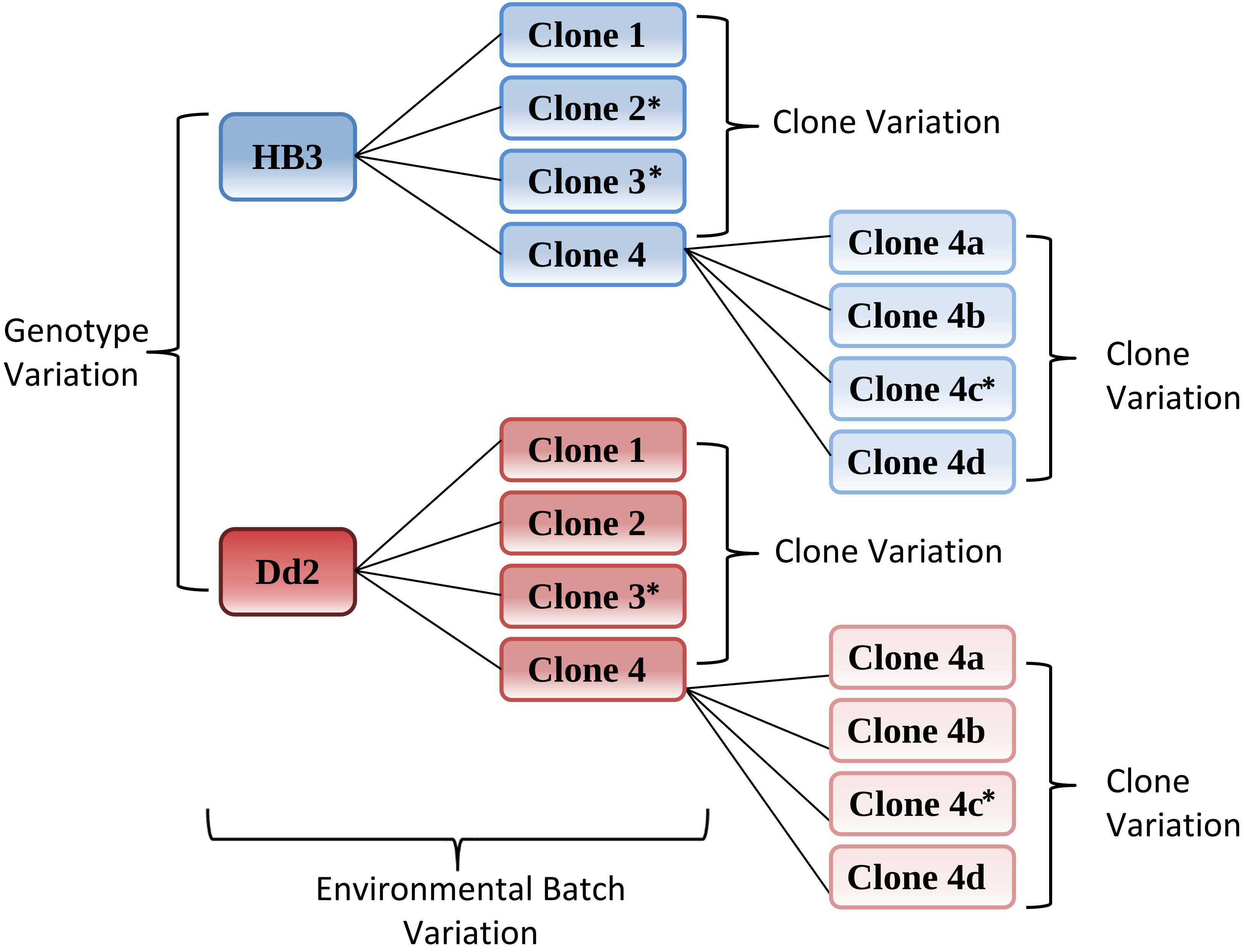
**Fig 2.** Correlations to reference lifecycle time points confirms parasite developmental stages. Whole genome normalized transcription abundance values for each sample (*y*-axis) were correlated to hourly transcription profiles of the community standard 3D7 parasite strain (*x*-axis). Color ramp represents correlation values by sample with highest values in red and lowest in green. The hours post invasion (hpi) value corresponding to the highest correlation value for each sample was recorded and used to determine stage-based variation in transcription among the samples.

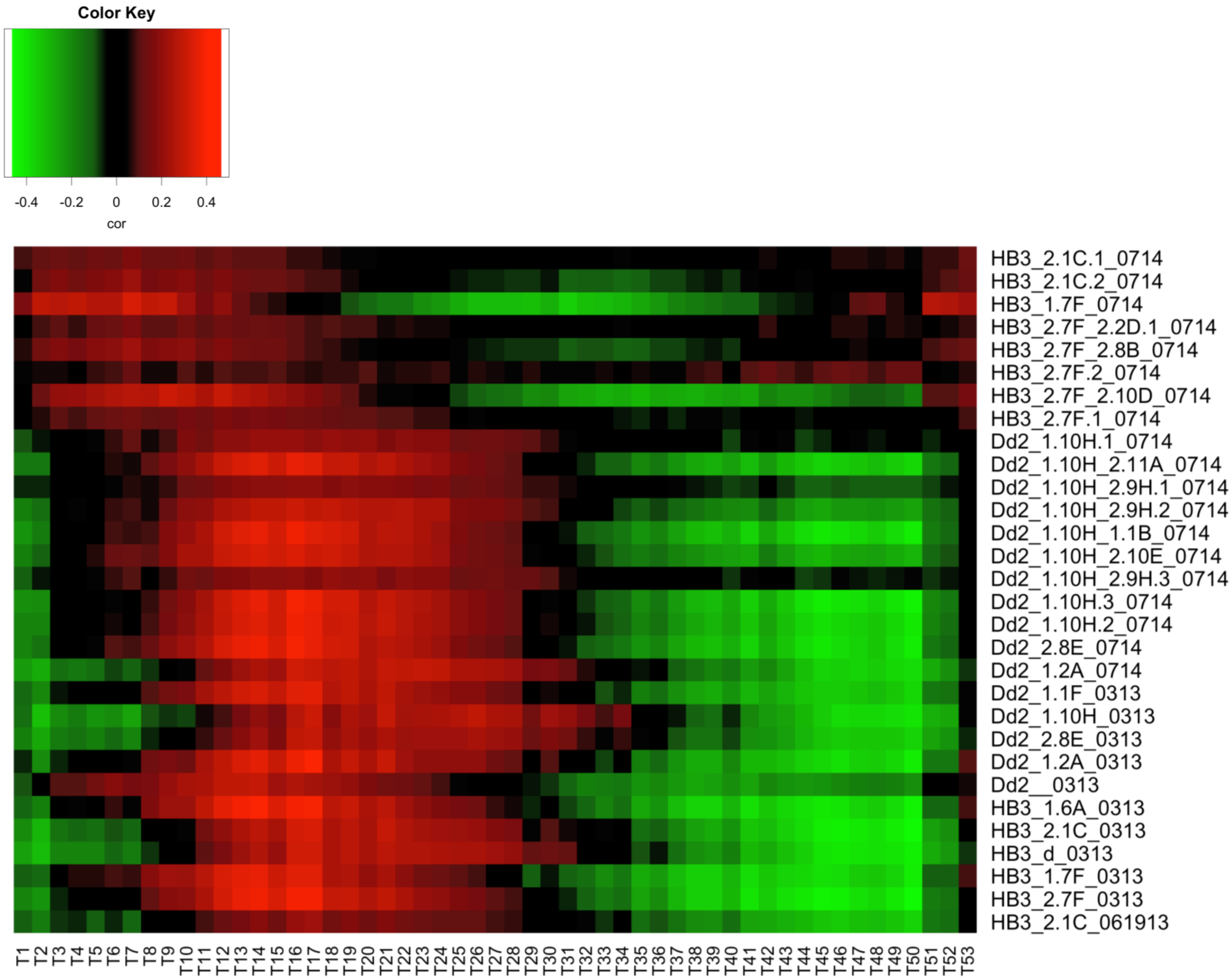
**Fig 3.** Partitioning of gene expression variation using a mixed effects model. Total variance in expression of genes PF3D7_0831700 (**A–E**) and PF3D7_1415800 (**F–J**) were sequentially partitioned (variation is removed stepwise) among identified sources. **A** and **F**: All variation is exhibited in the first panels. **B** and **G**: Stage-based variation is removed first and has little impact

on the total variation for these two genes. More variation between Dd2 and HB3 is evidenced in PF3D7_0831700 (**B**) by the difference in residuals between clustering points by genotype. **C** and **H**: After accounting for genotype, the scale of the residuals (unaccounted for expression variation) among the samples was significantly reduced in PF3D7_0831700 (**C**), but not in PF3D7_1415800 (**H**) in which a broad range of residuals still exists for both HB3 and Dd2. **D** and **I**: Removing genotype and environmental batch -based variation did not alter the overall residual variation for PF3D7_0831700 (**D**), but the range of residuals was significantly narrower for PF3D7_1415800 (**I**). **E** and **J**: Removing genotype, environmental batch, and clone-based variation did not significantly alter the final range in the residuals and the remaining variation among samples in panels (**E**) and (**J**) resulted from (unaccounted for) stochastic noise. Symbols indicate different samples: black circles for parental line, different colored circles for 1st round clones (filled for batch 1, outlined for batch 2), 2nd round clones are all red outlines with different shape for each unique clone. Residuals contain all total gene expression variation, across genotype, environmental batch, and clone.

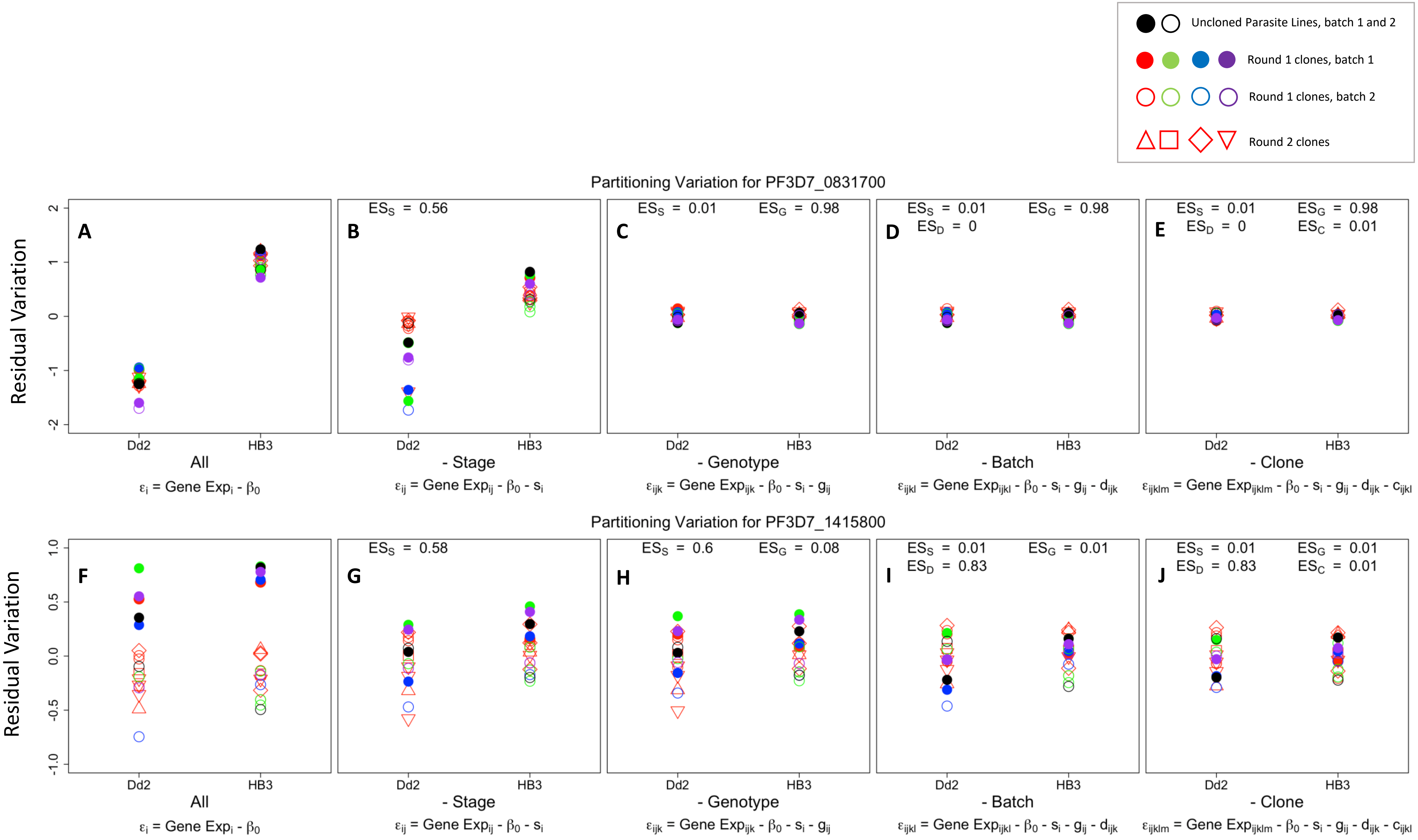
**Fig 4.** Significant transcriptional variation across all genes was observed for all the biological sources identified. **A**: The full multi-level mixed effects model partitioned the total variance for all 5540 genes and 100 noncoding RNAs among stage, genotype, environmental batch, and clone (Equation 9). **B**: For the 646 genes with significant variance, effect size for each source was also calculated. Each data point represents the transcriptional variance from one gene among all samples based on source. All genes are included for each source of variation. Boxes show inter-quartile distance (IQD) with a line centered on median values. Whiskers extend to 1.5 IQD. Genes with highly significant amounts of variance by source are represented as outlier points.

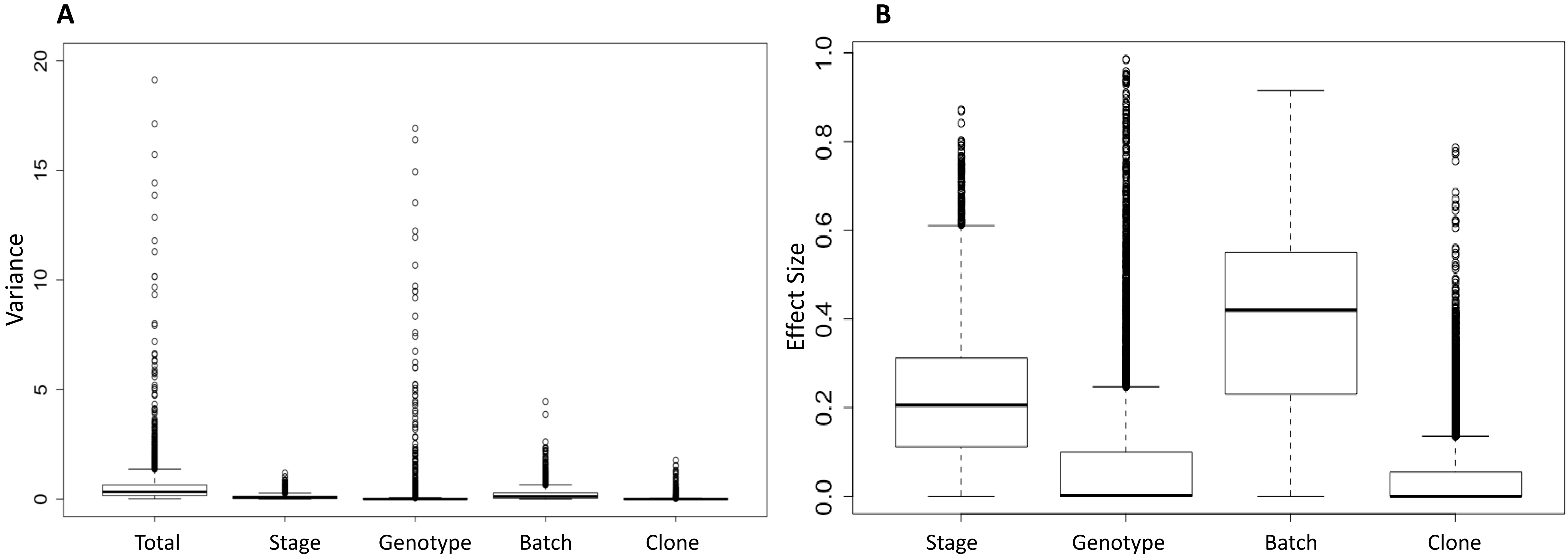
**Fig 5.** Transcripts largely exhibited significant variation for only once source. A total of 646 genes (*y*-axis) had significant variation based on one of the identified sources. For these genes most had significant levels of variation based on only a single source (yellow, $P < 0.05$, Bonferroni adjusted). Hierarchical clustering of significance values for each source of variation demonstrated that most (472) of these genes varied based on environmental batch, with another 171 varying for genotype, and 12 for clone. Only nine genes were significant for multiple sources, seven for genotype and environmental batch, and two for genotype and clone. Pairing the effect size directly with the total variance and significance demonstrates that even genes with low overall transcription variance can have significant variation due to a single source.


HB3
Clone 1
Clone 2*
Clone 3*
Clone 4
Clone Variation
Clone 4a
Clone 4b
Clone 4c*
Clone 4d
Clone Variation
Genotype Variation
Dd2
Clone 1
Clone 2
Clone 3*
Clone 4
Clone Variation
Clone 4a
Clone 4b
Clone 4c*
Clone 4d
Clone Variation
Environmental Batch Variation

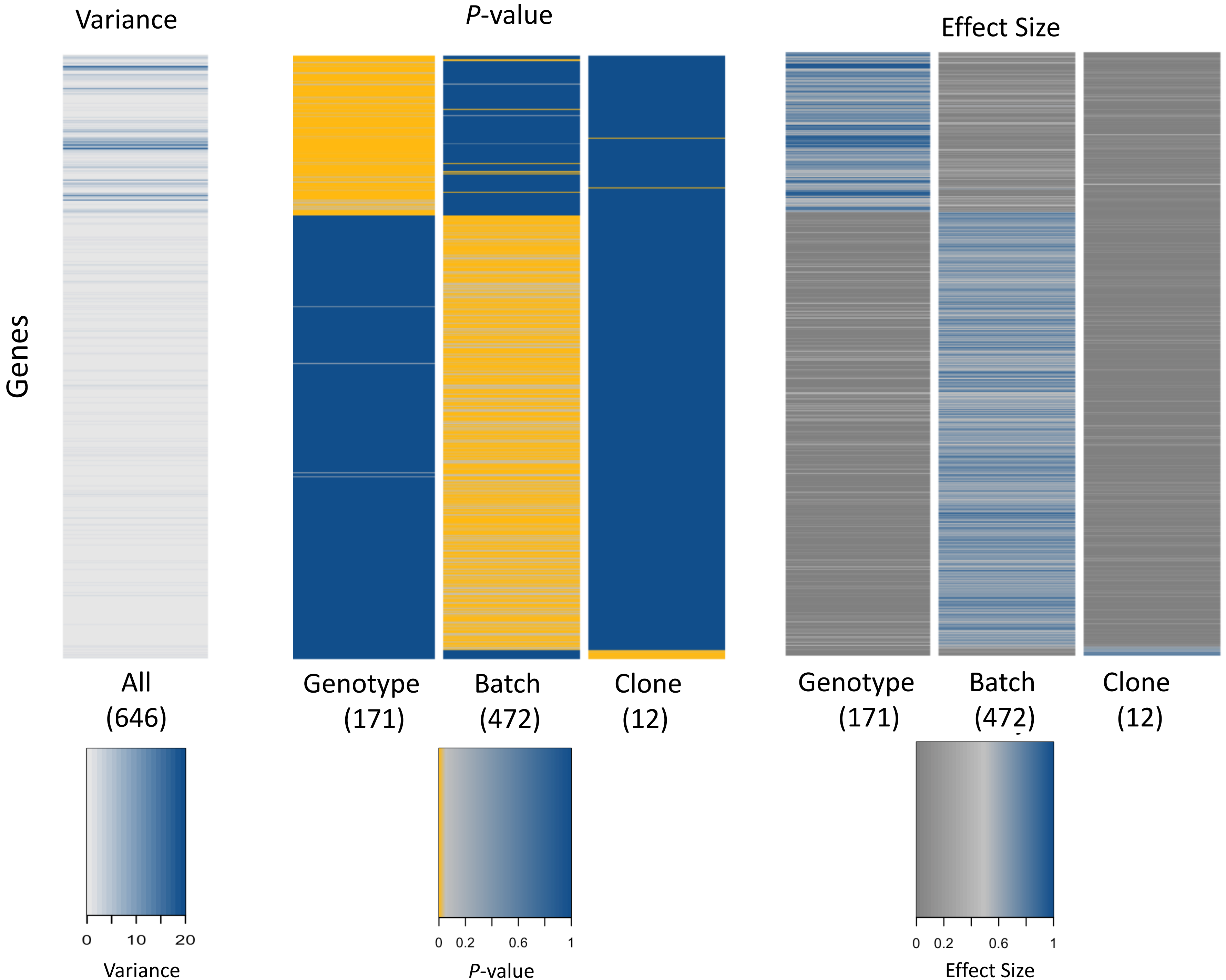


Color Key
-0.4
-0.2
0
0.2
0.4
cor
HB3_2.1C.1_0714
HB3_2.1C.2_0714
HB3_1.7F_0714
HB3_2.7F_2.2D.1_0714
HB3_2.7F_2.8B_0714
HB3_2.7F.2_0714
HB3_2.7F_2.10D_0714
HB3_2.7F.1_0714
Dd2_1.10H.1_0714
Dd2_1.10H_2.11A_0714
Dd2_1.10H_2.9H.1_0714
Dd2_1.10H_2.9H.2_0714
Dd2_1.10H_1.1B_0714
Dd2_1.10H_2.10E_0714
Dd2_1.10H_2.9H.3_0714
Dd2_1.10H.3_0714
Dd2_1.10H.2_0714
Dd2_2.8E_0714
Dd2_1.2A_0714
Dd2_1.1F_0313
Dd2_1.10H_0313
Dd2_2.8E_0313
Dd2_1.2A_0313
Dd2__0313
HB3_1.6A_0313
HB3_2.1C_0313
HB3_d_0313
HB3_1.7F_0313
HB3_2.7F_0313
HB3_2.1C_061913
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
T15
T16
T17
T18
T19
T20
T21
T22
T23
T24
T25
T26
T27
T28
T29
T30
T31
T32
T33
T34
T35
T36
T37
T38
T39
T40
T41
T42
T43
T44
T45
T46
T47
T48
T49
T50
T51
T52
T53


Uncloned Parasite Lines, batch 1 and 2
Round 1 clones, batch 1
Round 1 clones, batch 2
Round 2 clones
Partitioning Variation for PF3D7_0831700
Residual Variation
2
1
0
-1
-2
A
B
$ES_S$ = 0.56
C
$ES_S$ = 0.01
$ES_G$ = 0.98
D
$ES_S$ = 0.01
$ES_G$ = 0.98
$ES_D$ = 0
E
$ES_S$ = 0.01
$ES_G$ = 0.98
$ES_D$ = 0
$ES_C$ = 0.01
Dd2
HB3
All
$\varepsilon_i$ = Gene $Exp_i - \beta_0$
- Stage
$\varepsilon_{ij}$ = Gene $Exp_{ij} - \beta_0 - s_i$
- Genotype
$\varepsilon_{ijk}$ = Gene $Exp_{ijk} - \beta_0 - s_i - g_{ij}$
- Batch
$\varepsilon_{ijkl}$ = Gene $Exp_{ijkl} - \beta_0 - s_i - g_{ij} - d_{ijk}$
- Clone
$\varepsilon_{ijklm}$ = Gene $Exp_{ijklm} - \beta_0 - s_i - g_{ij} - d_{ijk} - c_{ijkl}$
Partitioning Variation for PF3D7_1415800
Residual Variation
1.0
0.5
0.0
-0.5
-1.0
F
G
$ES_S$ = 0.58
H
$ES_S$ = 0.6
$ES_G$ = 0.08
I
$ES_S$ = 0.01
$ES_G$ = 0.01
$ES_D$ = 0.83
J
$ES_S$ = 0.01
$ES_G$ = 0.01
$ES_D$ = 0.83
$ES_C$ = 0.01
Dd2
HB3
All
$\varepsilon_i$ = Gene $Exp_i - \beta_0$
- Stage
$\varepsilon_{ij}$ = Gene $Exp_{ij} - \beta_0 - s_i$
- Genotype
$\varepsilon_{ijk}$ = Gene $Exp_{ijk} - \beta_0 - s_i - g_{ij}$
- Batch
$\varepsilon_{ijkl}$ = Gene $Exp_{ijkl} - \beta_0 - s_i - g_{ij} - d_{ijk}$
- Clone
$\varepsilon_{ijklm}$ = Gene $Exp_{ijklm} - \beta_0 - s_i - g_{ij} - d_{ijk} - c_{ijkl}$



A
Variance
0
5
10
15
20
Total
Stage
Genotype
Batch
Clone
B
Effect Size
0.0
0.2
0.4
0.6
0.8
1.0
Stage
Genotype
Batch
Clone


Variance
*P*-value
Effect Size
Genes
All
(646)
Genotype
(171)
Batch
(472)
Clone
(12)
Genotype
(171)
Batch
(472)
Clone
(12)
0
10
20
Variance
0
0.2
0.6
1
*P*-value
0
0.2
0.6
1
Effect Size