Edinburgh Research Explorer

Genome annotation improvements from cross-phyla proteogenomics and time-of-day differences in malaria mosquito proteins using untargeted quantitative proteomics

Citation for published version:
https://doi.org/10.1371/journal.pone.0220225

Digital Object Identifier (DOI):
10.1371/journal.pone.0220225

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
PLoS ONE

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
RESEARCH ARTICLE

Genome annotation improvements from cross-phyla proteogenomics and time-of-day differences in malaria mosquito proteins using untargeted quantitative proteomics

Lisa Imrie1*, Thierry Le Bihan1,2,3*, Aíne O'Toole4, Paul V. Hickner5, W. Augustine Dunn6, Benjamin Weise2, Samuel S. C. Rund2,5*

1 SynthSys–Synthetic and Systems Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, 2 Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, United Kingdom, 3 Rapid Novor, Kitchener, Ontario, Canada, 4 Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom, 5 Eck Institute for Global Health, University of Notre Dame, Notre Dame, Indiana, United States of America, 6 Boston Children’s Hospital, Boston, Massachusetts, United States of America

* These authors contributed equally to this work.
* srund@nd.edu

Abstract

The malaria mosquito, *Anopheles stephensi*, and other mosquitoes modulate their biology to match the time-of-day. In the present work, we used a non-hypothesis driven approach (untargeted proteomics) to identify proteins in mosquito tissue, and then quantified the relative abundance of the identified proteins from *An. stephensi* bodies. Using these quantified protein levels, we then analyzed the data for proteins that were only detectable at certain times-of-the day, highlighting the need to consider time-of-day in experimental design. Further, we extended our time-of-day analysis to look for proteins which cycle in a rhythmic 24-hour (“circadian”) manner, identifying 31 rhythmic proteins. Finally, to maximize the utility of our data, we performed a proteogenomic analysis to improve the genome annotation of *An. stephensi*. We compare peptides that were detected using mass spectrometry but are ‘missing’ from the *An. stephensi* predicted proteome, to reference proteomes from 38 other primarily human disease vector species. We found 239 such peptide matches and reveal that genome annotation can be improved using proteogenomic analysis from taxonomically diverse reference proteomes. Examination of ‘missing’ peptides revealed reading frame errors, errors in gene-calling, overlapping gene models, and suspected gaps in the genome assembly.

Introduction

*Anopheles stephensi* is a major malaria vector in southern Asia where its geographic range extends across the Indian subcontinent [1]. Research on the African *Anopheles gambiae* mosquito has demonstrated that the behavior and physiology of the mosquito is highly dependent on circadian biology and time-of-day. For example, ~20% of *An. gambiae* genes were
The specific roles of all authors are articulated in the 'author contributions' section.

Competing interests: TLB has received salary from Rapid Novor. TLB’s employment at Rapid Novor does not alter our adherence to PLOS ONE policies on sharing data and materials. The other authors have declared that no competing interests exist.

The LC-MS QExactive equipment was purchased by a Wellcome Trust Institutional Strategic Support Fund and a strategic award from the Wellcome Trust for the Centre for Immunity, Infection and Evolution (095831/Z/11/Z). Rapid Novor provided support in the form of a salary for author TLB, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The remaining funders also had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Anopheles gambiae mosquito antennae utilized targeted quantitative proteomics, in which the mass spectrometer was tuned to specifically identify and quantify the protein abundance of proteins from an _a priori_ list of genes of interest [3] where only targeted proteins are interrogated. Targeted proteomics is a powerful technique, allowing the verification of a defined working hypothesis on specific proteins that are quantified.

In the present work, we used a non-hypothesis driven approach (untargeted proteomics) to identify proteins in mosquito tissue. In addition, we quantified the relative abundance of the identified proteins from _An. stephensi_ bodies. Such an untargeted, label free, quantitative analysis has been used on diverse tissues such as mammalian cells, yeast, bacteria, and _Ostreococcus tauri_ algae [23–26]. Using these quantified protein levels, we then analyzed the data for proteins that were only detectable at certain times-of-the-day, highlighting the need to consider time-of-day in experimental design. Further, we extended our time-of-day analysis to look for genes that are not only detectable at certain times-of-day, but which cycle in a rhythmic 24-hour ("circadian") manner.

Annotation of the _An. stephensi_ genome is far less complete than that of the model mosquito, _An. gambiae_. Proteogenomic analysis can be used to improve these annotations, particularly by experimentally validating computationally-derived open reading frame (ORF) predictions [27]. Additionally, proteogenomic analysis can be used in the identification of variant sequences and novel splicing sites [28, 29]. In these analyses, peptides detected using mass spectrometry were compared with reference proteomes. Peptide sequences not found in predicted protein coding regions indicate mis-annotations such as missing exons or entire missing genes, while peptide sequences with single amino-acid differences between the experimentally detected sequence and reference proteome may represent sequencing errors or polymorphisms. Here we apply proteogenomic analysis to _An. stephensi_, but employing a novel protocol, we reference over 30 predicted proteomes from other vector species to search against our experimentally derived peptides. We reveal that genome annotation can be improved using proteogenomic analysis from taxonomically diverse reference proteomes.
Results and discussion

Global survey of the proteome without fractionation

Using an untargeted proteomics approach without any fractionation on pooled samples of 10 whole mosquitoes harvested across the 24hr day, we have identified 12641 unique peptides (having a Maxquant score of 45 and more) mapping to ~1700 (identified with at least 2 peptides) An. stephensi proteins (S1 Table). An. stephensi is thought to have ~11,789 genes [9] thus, with no fractionation we observed 13% of the predicted proteome. Future studies could increase the number of detectable proteins by fractionation or by using an iTRAQ or TMT labelling strategy combined with fractionation as a good compromise between high number of samples and fractionation.

Time-of-day dependent changes in detectable genes

In order to determine if there is a time-of-day dependent ability to detect proteins, mosquito bodies were collected every four hours from three staggered time courses (Fig 1) of between 28–44 hr (Fig 1). Note in this collection protocol, each of five times-of-day are sampled, independently, five times (see Fig 1). Untargeted, quantitative proteomics was performed (S2 Table, S3 Table), and a total of 1525 body proteins were deemed quantifiable (identified with a least two component peptides) from the body samples.

Our data revealed differences in the total number of proteins that were quantifiable at any given time-of-day (583–733 proteins per time-of-day) (Fig 2A). When the identity of proteins detectable (i.e. quantity > 0) is considered, there are a number of proteins only detectable at certain times-of-day. Whereas 489 proteins were detectable at all sampling times, there were between 45 and 72 proteins that were only detectable at 2–5 times-of-day, and 134 were only detectable at a single time-of-day (Fig 2B). Not surprisingly, proteins that had a higher average abundance were generally detectable at more times-of-day (Fig 2C).

Identification of rhythmic genes

We extended our analysis further by looking for genes that were rhythmically expressed using an algorithm (JTK_CYCLE) specifically designed for looking for “circadian” expression patterns. First, we analyzed the subset of 357 proteins that were detected at all time points, and where >1 peptide was used to identify each protein in each sample at each timepoint. Of these
proteins, ANOVA revealed 90 proteins, where at least one time point was significantly different from the others \((p < 0.1)\). Next, from the list of proteins with statistically significant time-of-day differences in protein concentration (90 proteins), we proceeded to analyze those proteins for 24 hr daily rhythms in abundance (rather than only a simple time-of-day difference) using the JTK_CYCLE algorithm. This algorithm is used to mine ‘omic data for such 24 hour rhythms \([30, 31]\), and we thus applied it to our data. JTK_CYCLE identified 31 proteins as having rhythmic expression (Fig 3). As \(A.\) stephensi has not been extensively annotated, we mapped these 31 proteins to their homologues in \(Ae\). aegypti, \(An\). gambiae, \(Culex\) quinquefasciatus, and/or \(D\). melanogaster to assign a name/function to each protein (Table 1). We note these rhythmic proteins display a wide range of phases (times-of-day when proteins peak) and abundance amplitudes (Fig 3, Table 1).

**Conservation of rhythmicity across species.** We next searched published studies of rhythmic gene expression in \(Ae\). aegypti \([32]\), \(An\). gambiae \([2]\), and \(D\). melanogaster \([33]\) to determine if homologues of rhythmic \(A.\) stephensi protein were rhythmic in these species at the gene expression level. Indeed, we determined that of our 31 rhythmically identified \(A.\) stephensi proteins, 17 had homologues in at least one of the other three species with rhythmic expression of the same gene (Table 1). This represents ~55% of identified \(A.\) stephensi rhythmic proteins. For example, considering \(A.\) stephensi protein ASTE101494, which is predicted to be glycerol 3-phosphate based on homology to \(D.\) melanogaster (FBgn0001128) and \(An\). gambiae (AGAP007593), we find the \(A.\) stephensi protein abundance is rhythmic, as are the gene expression levels in \(An\). gambiae gene expression and \(D.\) melanogaster expression levels. Similarly, protein abundance levels of ASTE100584 and expression levels of the homologous hydrogenase genes in \(Ae\). aegypti (AAEL010814)/\(An\). gambiae (AGAP003167) are also both rhythmic (see Table 1).

**Proteogenomic analysis.** In order to determine if our proteomics work could be used to improve the \(A.\) stephensi genome, we next performed a proteogenomic analysis. Two sets of
computed proteomes from VectorBase were utilized: (1) the *An. stephensi* proteome (Indian strain peptide sequences, Astel2.3 geneset with 11,789 entries); and (2) the complete proteomes stored in VectorBase (here referred to as "All Vectors", with $>566,000$ entries). "All vectors" comprises 39 proteomes (S4 Table), including other mosquito proteomes and other
Table 1. Rhythmic proteins identified.

<table>
<thead>
<tr>
<th>Indian-strain geneID</th>
<th>SDA-500 geneID</th>
<th>JTK_CYCLE values</th>
<th>Dipteran homologues **</th>
<th>Ensembl ID</th>
<th>Species: Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASTE00121</td>
<td>ASTE003060</td>
<td>0.013 18</td>
<td>CPIJ013361</td>
<td>CQUI: Tropomyosin-1</td>
<td></td>
</tr>
<tr>
<td>ASTE00584</td>
<td>ASTE004515</td>
<td>0.018 20</td>
<td>AAEL010814</td>
<td>AGAM: isocitrate dehydrogenase</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ASTE008899</td>
<td></td>
<td>AGAP003167</td>
<td>AGAM: NAD(P) transhydrogenase</td>
<td></td>
</tr>
<tr>
<td>ASTE00675</td>
<td>ASTE008899</td>
<td>0.015 8</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASTE00949</td>
<td>ASTE002146</td>
<td>0.090 16</td>
<td>AGAP003937</td>
<td>AGAM: AGAP003937</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ASTE011453</td>
<td></td>
<td>AGAP003936</td>
<td>AGAM: small nuclear ribonucleoprotein D2</td>
<td></td>
</tr>
<tr>
<td>ASTE01125</td>
<td>ASTE005445</td>
<td>0.030 20</td>
<td>CPIJ011528</td>
<td>CQUI: NADH dehydrogenase iron-sulfur protein 2, mitochondrial</td>
<td></td>
</tr>
<tr>
<td>ASTE01494</td>
<td>ASTE001799</td>
<td>0.012 22</td>
<td>FBgn0019968</td>
<td>DMEL: Kinesin-73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ASTE001800</td>
<td></td>
<td>FBgn0001128</td>
<td>DMEL: GPD-C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ASTE001801</td>
<td></td>
<td>AGAP007593</td>
<td>AGAM: glycerol-3-phosphate dehydrogenase (NAD+)</td>
<td></td>
</tr>
<tr>
<td>ASTE01887</td>
<td>ASTE007267</td>
<td>0.020 8</td>
<td>AAEL003211</td>
<td>AAEG: beta-carotene dioxygenase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGAP008143</td>
<td>AGAM: AGAP008143</td>
<td></td>
</tr>
<tr>
<td>ASTE02163</td>
<td>ASTE010238</td>
<td>0.063 20</td>
<td>AGAP005558</td>
<td>AGAM: peptidase (mitochondrial processing) beta</td>
<td></td>
</tr>
<tr>
<td>ASTE02221</td>
<td>ASTE009536</td>
<td>0.026 20</td>
<td>AGAP005627</td>
<td>AGAM: creatine kinase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGAP012924</td>
<td>AGAM: Arginine kinase</td>
<td></td>
</tr>
<tr>
<td>ASTE02598</td>
<td>ASTE008307</td>
<td>0.056 18</td>
<td>AGAP006936</td>
<td>AGAM: Mitochondrial cytochrome c1 heme protein</td>
<td></td>
</tr>
<tr>
<td>ASTE02810</td>
<td>ASTE010498</td>
<td>0.071 18</td>
<td>CPIJ000098</td>
<td>CQUI: Electron transfer flavoprotein-ubiquinone oxidoreductase</td>
<td></td>
</tr>
<tr>
<td>ASTE03561</td>
<td>ASTE000276</td>
<td>0.001 10</td>
<td>FBgn0263594</td>
<td>DMEL: RH35990p</td>
<td></td>
</tr>
<tr>
<td>ASTE03834</td>
<td>ASTE011384</td>
<td>0.056 4</td>
<td>AAEL007698</td>
<td>AAEG: PIWI</td>
<td></td>
</tr>
<tr>
<td>ASTE03928</td>
<td>ASTE003184</td>
<td>0.001 14</td>
<td>AGAP010051</td>
<td>AGAM: AGAP010051</td>
<td></td>
</tr>
<tr>
<td>ASTE04436</td>
<td>ASTE006286</td>
<td>0.090 18</td>
<td>AGAP000720</td>
<td>AGAM: Neuronal cell adhesion molecule</td>
<td></td>
</tr>
<tr>
<td>ASTE05481</td>
<td>ASTE011491</td>
<td>0.044 16</td>
<td>CPIJ019398</td>
<td>CQUI: Myosin light chain 2</td>
<td></td>
</tr>
<tr>
<td>ASTE05637</td>
<td>ASTE010982</td>
<td>0.009 18</td>
<td>AGAP011131</td>
<td>AGAM: F-type H+-transporting ATPase subunit d</td>
<td></td>
</tr>
<tr>
<td>ASTE06078</td>
<td>ASTE002102</td>
<td>0.050 22</td>
<td>AGAP012100</td>
<td>AGAM: 40S ribosomal protein S26</td>
<td></td>
</tr>
<tr>
<td>ASTE06447</td>
<td>ASTE009294</td>
<td>0.003 2</td>
<td>AAEL003427</td>
<td>AAEG: 40S ribosomal protein S16</td>
<td></td>
</tr>
<tr>
<td>ASTE06644</td>
<td>ASTE008693</td>
<td>0.008 8</td>
<td>AGAP008364</td>
<td>AGAM: thioester-containing protein 15</td>
<td></td>
</tr>
<tr>
<td>ASTE06854</td>
<td>ASTE001249</td>
<td>0.034 10</td>
<td>FBgn0031021</td>
<td>DMEL: NADH dehydrogenase (ubiquinone) 18 kDa subunit</td>
<td></td>
</tr>
<tr>
<td>ASTE07075</td>
<td>ASTE002586</td>
<td>0.026 12</td>
<td>AGAP004055</td>
<td>AGAM: 2-oxoglutarate dehydrogenase E2 component</td>
<td></td>
</tr>
<tr>
<td>ASTE07469</td>
<td>ASTE001425</td>
<td>0.071 18</td>
<td>AAEL014913</td>
<td>AAEG: Pyruvate kinase</td>
<td></td>
</tr>
<tr>
<td>ASTE075885</td>
<td>ASTE010928</td>
<td>0.071 10</td>
<td>AGAP004146</td>
<td>AGAM: Ras-related protein Rab-1A</td>
<td></td>
</tr>
<tr>
<td>ASTE08990</td>
<td>ASTE000090</td>
<td>0.007 10</td>
<td>AGAP010895</td>
<td>AGAM: spectrin beta</td>
<td></td>
</tr>
<tr>
<td>ASTE09101</td>
<td>ASTE006195</td>
<td>0.003 18</td>
<td>AGAP007841</td>
<td>AGAM: F-type H+-transporting ATPase subunit delta</td>
<td></td>
</tr>
<tr>
<td>ASTE09205</td>
<td>ASTE008202</td>
<td>0.023 18</td>
<td>AGAP011800</td>
<td>AGAM: Transaldolase</td>
<td></td>
</tr>
<tr>
<td>ASTE09484</td>
<td>ASTE004410</td>
<td>0.034 18</td>
<td>AGAP010404</td>
<td>AGAM: Glutathione S-transferase</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AAEL011741</td>
<td>AAEG: AAEL011741</td>
<td></td>
</tr>
<tr>
<td>ASTE09679</td>
<td>ASTE002472</td>
<td>0.020 10</td>
<td>AAEL007881</td>
<td>AAEG: AAEL007881</td>
<td></td>
</tr>
<tr>
<td>ASTE110862</td>
<td>ASTE001429</td>
<td>0.004 10</td>
<td>FBgn0016693</td>
<td>DMEL: Putative Achaete Scute Target 1</td>
<td></td>
</tr>
<tr>
<td>ASTE11367</td>
<td>No match</td>
<td>0.012 16</td>
<td>AGAP007122</td>
<td>AGAM: Tubulin, alpha 1</td>
<td></td>
</tr>
</tbody>
</table>

* The listed homologue has been found to be expressed rhythmically in *Aedes aegypti* [32], *Anopheles gambiae* [2], or *Drosophila melanogaster* [33]

** Only rhythmic homologues and/or a representative named homologue in AAEG: *Ae. aegypti*, AGAM: *An. gambiae*, CQUI: *Culex quinquefasciatus*, or DMEL: *D. melanogaster* is provided.

*** The calculated time-of-day, in zeitgeber (ZT) time, when protein abundance peaks where ZT0 is lights on and ZT12 is lights off.

https://doi.org/10.1371/journal.pone.0220225.t001
vectors such as snails, ticks, and kissing bugs. In other words, we analyzed the spectra generated from each peptide and compared it against both a list of *An. stephensi* computed peptide spectra and our “All Vectors” computed peptide spectra. A spectrum that was found in “All Vectors” but not *An. stephensi* was deemed to be missing from the *An. stephensi* genome, since we had proteomic evidence of the existence of the peptide as computed from other species (of varying degrees of relatedness).

There were 57,726 peptide matches between our data and the *An. stephensi* (Indian strain) proteome in VectorBase. Using a score cut-off based on identity score (average of 35.6 for “All Vectors” and 19 for *An. stephensi*) the number of retained high confident peptides matches decreased to 82,402 and 32,949 for “All Vectors” and *An. stephensi*, respectively. We manually validated all hits which were found in both sets and removed any peptides that had identical or nearly identical scores in both “All Vectors” and *An. stephensi* (e.g. isoleucine to leucine mutations are isobaric and are thus not easily distinguishable by mass spectrometry).

From the preceding analyses and filtering (such as excluding matches to SDA-500, another *An. stephensi* strain) we identified a total of 792 (S5 Table) high-confidence matches between peptide sequences that are ‘missing’ from the *An. stephensi* (Indian strain) proteome, yet are found in the 39 “All Vectors” (S4 Table) proteomes. Some identified ‘missing’ peptides had only a match in another single organism, while other peptides could be found in up to 35 other proteomes. By also combining peptide sequences that are completely contained within longer peptide sequences that were also detected, we are left with 239 unique peptide groups (e.g. ATAQLIESIK, ATAQLIE, and ATAQL count as one peptide group) that were detected using proteomics in *An. stephensi* samples, that matched at least one peptide sequence in one of the 39 other proteomes, but are not found in the currently available predicted Indian strain *An. stephensi* proteome. Matches were found across a wide phylogenetic diversity, from lice to snails—not just in other mosquitoes or diptera (Supplemental 5, Fig 4). These matches suggest possible genome annotation errors, which we next analyzed.

**Identification of potential errors in the genome annotations.** Further analysis of the peptides detected in our study, but missing from the *An. stephensi* proteome (Indian strain, Astel2.3), revealed potential errors in the genome annotations and/or assembly (S6 Table). Of the 239 peptides missing from the *An. stephensi*, but with matches found in “All Vectors,” were 2 peptides with 100% identity to a transcript based on tBLASTn analysis. These were found to be in a different reading frame than the annotated transcript and represent missing gene models where two genes overlap—a common phenomenon in eukaryotic organisms [34]. tBLASTn analysis revealed 25 peptides with 100% identity to a genome scaffold but not to a transcript. These are most likely genes that were not called by the gene prediction software and are missing from the current geneset (Astel2.3). There were 94 peptides that had high homology to a genomic region that contained a SNP or a mismatch causing a frameshift mutation, which could be either mutations or sequencing errors. Finally, 120 peptides were not found in the genome using standard BLAST tools and may represent gaps in the genome assembly.

**Conclusions**

In this work we performed untargeted quantitative proteomics on *An. stephensi* mosquito samples to answer three different questions: (1) Are there qualitative, time-of-day differences in the peptides? (2) What proteins can be detected as rhythmic in a 24 hour “circadian” manner? and (3) Can the *An. stephensi* genome be improved using proteomic data compared against genomes of other species?
By collecting mosquito samples every four hours across the day, we determined that there are time-of-day differences in the number and quantity of proteins that are detectable at any given time-of-day. Previous work in *An. gambiae* revealed dusk (ZT12) was the time-of-day that had the greatest number of rhythmic genes which had their peak in expression [2]. It was hypothesized that this is due to the massive change in mosquito behavior and physiology as it goes from a resting state during the day to an active, host seeking mosquito at night [2]. Congruent with this, the greatest number of detectable proteins were detected at dusk in *An. stephensi*. Not surprisingly, proteins that had a higher protein abundance were detectable at more times-of-the-day. A total of 134 proteins were detected only at one time-of-day, and 489 were detectable at all times-of-day.

We next extended our analysis from a question of detectability at different times-of-day, to see if 24-hour (“circadian”) levels of protein abundance could be detected in our dataset. Previous work in the *An. gambiae* mosquito utilized targeted quantitative proteomics, whereby the protein abundance of proteins from an *a priori* list of genes of interest [3, 17] was quantified. Here we attempted to use a non-hypothesis driven, untargeted proteomic approach to quantify proteins in mosquito tissue. Work in other species has previously revealed ~20% of *An. gambiae* genes and at least 8% *Ae. aegypti* genes are rhythmically expressed over the 24-hour day [2].

Fig 4. *An. stephensi* proteogenomic analysis revealed evidence for the presence of *An. stephensi* peptide sequences that are not in the predicted *An. stephensi* (Indian strain) proteome but match predicted peptide sequences from species across a wide-range of taxa. Numbers listed as matches represent the number of ‘missing’ *An. stephensi* peptide groups found in the given taxonomic group. *Anopheles* species excludes the *An. stephensi* (SDA-500) strain.

https://doi.org/10.1371/journal.pone.0220225.g004

By collecting mosquito samples every four hours across the day, we determined that there are time-of-day differences in the number and quantity of proteins that are detectable at any given time-of-day. Previous work in *An. gambiae* revealed dusk (ZT12) was the time-of-day that had the greatest number of rhythmic genes which had their peak in expression [2]. It was hypothesized that this is due to the massive change in mosquito behavior and physiology as it goes from a resting state during the day to an active, host seeking mosquito at night [2]. Congruent with this, the greatest number of detectable proteins were detected at dusk in *An. stephensi*. Not surprisingly, proteins that had a higher protein abundance were detectable at more times-of-the-day. A total of 134 proteins were detected only at one time-of-day, and 489 were detectable at all times-of-day.

We next extended our analysis from a question of detectability at different times-of-day, to see if 24-hour (“circadian”) levels of protein abundance could be detected in our dataset. Previous work in the *An. gambiae* mosquito utilized targeted quantitative proteomics, whereby the protein abundance of proteins from an *a priori* list of genes of interest [3, 17] was quantified. Here we attempted to use a non-hypothesis driven, untargeted proteomic approach to quantify proteins in mosquito tissue. Work in other species has previously revealed ~20% of *An. gambiae* genes and at least 8% *Ae. aegypti* genes are rhythmically expressed over the 24-hour day [2].
Here, in *An. stephensi*, we were able to detect rhythmic protein abundance levels in only 31 of the 1525 quantifiable proteins (2%). This is much lower than we expected from gene expression data from other mosquito species, but likely represents an underestimation of the true number of rhythmic proteins. One explanation for undetected rhythmic protein abundance is that low-concentration protein time courses were removed prior to analysis. These peptides may have represented protein abundance levels that were rhythmic, but dropped below the detection limits in our experimental runs at certain times-of-day. To our knowledge, this is the first untargeted quantitative proteomics performed in a mosquito species, and we reveal it can be used to reliably quantify a large number of proteins. We note, however, that some proteins may be rhythmic, or only appear at certain times-of-day, and this point should be considered when doing experimental design.

When we considered the homology of proteins here found to have rhythmic abundance levels, ~55% of identified *An. stephensi* rhythmic proteins have homologues in other species that have been determined, at the gene expression level, to be also rhythmic. This provides further evidence of the rhythmic nature of biological processes being conserved across species [35].

Finally, the present study revealed a number of potential errors in the current *An. stephensi* genome annotations and/or assembly. Untargeted proteomics could be leveraged to improve current genome annotations; however, proteomic reducibility, speed, and whole-proteome coverage are limited using our current technologies.

**Materials and methods**

**Biological material**

A lab colony of *An. stephensi* mosquitoes were maintained at ~60% relative humidity and 26˚C on a 12 hr/12 hr LD cycle [11 hr full light, 11 hr darkness (0 lux) and 1 hr dawn and 1 hr dusk transitions]. Access to 8% (w/v) fructose was provided *ad libitum*. In three replicate time courses (with slightly different durations depending on available number of mosquitoes per batch, see Fig 1), mosquitoes were placed in individual containers (pots) and allowed to acclimate for several days. A pot of mosquitoes was euthanized on dry ice every four hours and placed in -80˚C prior to tissue preparation. Heads were separated on dry ice from bodies (legs and wings were removed). Herein “body” describes the body of the mosquito with no head, wings, or legs.

**Sample preparation for the time series analysis**

A pool of 10 bodies was used per sample, solubilized in 250ul 8M urea 1% SDS and homogenized using a Precellys cell homogenizer (Bertin Instruments). The homogenization step comprises three steps of 40s at 5000rpm with a 10s pause; the overall procedure was repeated twice. A protein assay was performed using Pierce BCA protein assay kit, and a 50μg protein equivalent was used for SDS-PAGE analysis. Samples were briefly run on SDS-PAGE gel for 10 min, extracted and digested using Shevchenko’s method [36]. Peptide extracts were then cleaned on SPE reverse phase Bond Elut LMS (Agilent). The samples were dried under low pressure (Speedvac from Thermo-Fisher) and stored at -20˚C.

**HPLC-MS analysis**

The dried peptide samples were re-suspended in resuspension buffer (0.05%v/v trifluoroacetic acid in water) to a final concentration of 1 μg/μl. These samples were filtered using a Millex filter before subjecting to HPLC-MS analysis. Nano-HPLC-MS/MS analysis was performed using an on-line system consisting of a nano-pump (Dionex Ultimate 3000, Thermo-Fisher,
coupled to a QExactive instrument (Thermo-Fisher, UK) with a pre-column of 300 μm x 5 mm (Acclaim Pepmap, 5 μm particle size) connected to a column of 75 μm x 50 cm (Acclaim Pepmap, 3 μm particle size). Samples were analyzed on a 90 min gradient in data dependent analysis (1 survey scan at 70k resolution followed by the top 10 MS/MS).

Proteomics, protein identification and quantification
Data from MS/MS spectra were searched using MASCOT Versions 2.4 (Matrix Science Ltd, UK) against An. stephensi (Indian strain Astel2.3) data stored in VectorBase [37, 38]. Search parameters included a maximum missed-cut value set to 2. The following features were used in all searches: i) variable methionine oxidation, ii) fixed cysteine carbamidomethylation, iii) precursor mass tolerance of 10 ppm, iv) MS/MS tolerance of 0.05 amu, v) significance threshold (p) below 0.05 (MudPIT scoring) and vi) final Mascot peptide score of 20. A complete dataset was analysed using MaxQuant v1.5.2.8 [39] assuming a Maxquant score of 45 and more.

For the time series quantification analysis Progenesis (version 4, Nonlinear Dynamics) was used for LC-MS label-free quantitation (S2 Table). Progenesis QI for proteomics software has been designed specifically to perform label-free quantitation and is capable of analyzing significant numbers of large data files due to its peak-modelling algorithm which reduces the data files by an order of magnitude without losing any information. This allows for the analysis of large data sets including large numbers of replicates that would otherwise be impractical to run. The software is enabled with a graphical user interface which allows MS data to be viewed in either two or three dimensions. This can help to verify if features have been quantified accurately. In brief, the basic software steps are as follows: (1) Alignment of runs to compensate for LC separation "between-run" variation, allowing like-for-like comparison of peptide signals; (2) Feature detection and quantitation using peak area method; (3) Peptide identification using the mascot search engine; and (4) Peptide/protein quantitation using the calculated abundance of the features to which identifications have been matched.

Only MS/MS peaks with a charge of 2+, 3+ or 4+ were considered for the total number of ‘Features’ (signal at one particular retention time and m/z) and only the five most intense spectra per ‘Feature’ were included. Normalization was first performed based on the median of the ion intensities of these sets of multi-charged ions (2+, 3+, and 4+). The associated unique peptide ion intensities for a specific protein were then summed to generate an abundance value, which was transformed using an ArcSinH function (a log transform is not ideal considering the significant amount of near zero measurements generated by the current method of detection). Based on the abundance values, within group means were calculated and from there the fold changes (in comparison to control) were evaluated. One-way ANOVA was used to calculate the p-value based on the transformed values. A larger dataset with samples from both heads and bodies were analyzed using Maxquant to generate a list of identified protein and peptides (S3 Table). False Discovery Rate (FDR) information is provided in Table 2.

<table>
<thead>
<tr>
<th>Peptide matches above identity threshold</th>
<th>An. stephensi</th>
<th>Decoy</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide matches above homology or identity threshold</td>
<td>13300</td>
<td>248</td>
<td>1.86%</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0220225.t002

Table 2. False Discovery Rate (FDR) information.
**Homologue identification**

To generate the gene names and orthologues in Table 1, the list of strain Indian (ASTEI) target proteins detected in the time course proteomics analysis translations was matched to strain SDA-500 (ASTE) homologues. Using this translation table, Indian strain proteins were matched to dipteran orthologous genes in OrthoDB [40, 41]. (ftp://cegg.unige.ch/OrthoDB8/Eukaryotes/Genes_to_Ogs/ODB8_EukOGs_genes_ALL_levels.txt.gz). Filtering was performed in a Jupyter notebook [42] using the Pandas (v0.18.1) Python library [43]. The notebook (https://figshare.com/s/dbb89cb869f416979f60) is accessible on FigShare. Some gene names were manually supplemented using VectorBase [37, 38] or FlyBase [44].

**Statistical analysis for rhythmic genes**

In order to detect rhythmic protein abundance, we first only considered proteins where ANOVA revealed at least one time point is significantly different from the others (p < 0.1). Abundance data for those proteins was then processed with JTK_CYCLE [30] using the Meta-Cycle R package [31] to identify rhythmic proteins with a 24 hr period. Proteins were called rhythmic when their quantified protein abundance was determined by rhythmic by JTK_CYCLE (p < 0.1). We report at p = 0.09, q = 0.20 as our false discovery rate.

**Proteogenomic analysis**

A proteogenomic analysis was performed using the Mascot (Matrix science) package. Two sets of computed proteomes from VectorBase [37, 38] were utilized: (1) the *An. stephensi* proteome (Indian strain peptide sequences, AstI2.3 geneset with 11,789 entries); and (2) the entirety of arthropod proteomes stored in VectorBase (“All Vectors”, with >566,000 entries). This comprises 39 proteomes (S4 Table), both a second *An. stephensi* strain (SDA-500), other mosquito proteomes, snails, and other arthropod vectors such as sandflies, ticks, and kissing bugs.

We generated a subset of MS/MS features by removing MS/MS feature redundancy (keeping a maximum of the 5 most abundant peaks having the same masses and retention time). The merge .mgf file was generated using Progenesis. The dataset was searched against: (1) the *An. stephensi* proteome stored in VectorBase; and (2) against all the proteomes stored in VectorBase. We filter the identified peptide as follow: we only kept in both searches peptides having a Mascot score above identity and were ranked as first hit. We compared the two datasets and removed any specific peptide having the same score in both databases or showing similar peptide permutations such as isoleucine to leucine, which have indistinguishable mass spectrometry readings.

**Supporting information**

S1 Table. Identified peptides and proteins from untargeted proteomics.  
(XLSX)

S2 Table. Progenesis output.  
(XLSX)

S3 Table. Results of quantitative proteomics of *An. stephensi* bodies.  
(XLSX)

S4 Table. List of reference species and gene build versions from VectorBase used for proteogenomic analyses.  
(DOCX)
S5 Table. “Missing peptides” not found in *An. stephensi*, but with matches found in other species.

(CSV)

S6 Table. BLAST analysis of *An. stephensi* genome assembly (Astel2-Indian strain) using the peptide sequences from proteomics analysis revealed missing genome annotations. Peptide transcript hit (100%), 100% identity to *An. stephensi* gene transcript with full query length using tblastn; Scaffold hit (100%), 100% identity to *An. stephensi* genome scaffold with full query length using tblastn; Wobbly scaffold hit (<100% hit) <100% identity to *An. stephensi* genome scaffold with 1 mismatch and with full query length ± 1 with tblastn; No hit, no hits (not transcript, scaffold, or wobbly scaffold hit) in *An. stephensi* genome using tblastn, missing in genome assembly.

(CSV)

Acknowledgments

We thank Kimberly Prior, Aidan O’Donnell, and Ronnie Mooney for their assistance collecting and rearing mosquitoes during the time courses. We thank Sarah E. Reece for use of her lab space and equipment. We thank Michael Hughes, Gang Wu, and John Hogenesch for provision and assistance with JTK_CYCLE and MetaCycle. We thank Giles K.P. Bara for his useful comments on the manuscript.

Author Contributions

**Conceptualization:** Thierry Le Bihan, Samuel S. C. Rund.

**Formal analysis:** Thierry Le Bihan, Áine O’Toole, Paul V. Hickner, W. Augustine Dunn, Samuel S. C. Rund.

**Funding acquisition:** Samuel S. C. Rund.

**Investigation:** Lisa Imrie, Thierry Le Bihan, Benjamin Weise, Samuel S. C. Rund.

**Methodology:** Thierry Le Bihan.

**Supervision:** Samuel S. C. Rund.

**Writing – original draft:** Thierry Le Bihan, Áine O’Toole, Samuel S. C. Rund.

**Writing – review & editing:** Thierry Le Bihan, Paul V. Hickner, Samuel S. C. Rund.

References


