

# Identification of Cryptic *Anopheles* Mosquito Species by Molecular Protein Profiling

Pie Müller<sup>1,2\*</sup>, Valentin Pflüger<sup>3</sup>, Matthias Wittwer<sup>4</sup>, Dominik Ziegler<sup>3</sup>, Fabrice Chandre<sup>5</sup>, Frédéric Simard<sup>5</sup>, Christian Lengeler<sup>2,6</sup>

**1** Department of Medical Services and Diagnostic, Swiss Tropical and Public Health Institute, Basel, Switzerland, **2** University of Basel, Basel, Switzerland, **3** Mabritec AG, Riehen, Switzerland, **4** Spiez Laboratory, Swiss Federal Department of Defence, Spiez, Switzerland, **5** Maladies Infectieuses et Vecteurs: Ecologie, Génétique, Evolution et Contrôle, Institut de Recherche pour le Développement, Montpellier, France, **6** Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Basel, Switzerland

## Abstract

Vector control is the mainstay of malaria control programmes. Successful vector control profoundly relies on accurate information on the target mosquito populations in order to choose the most appropriate intervention for a given mosquito species and to monitor its impact. An impediment to identify mosquito species is the existence of morphologically identical sibling species that play different roles in the transmission of pathogens and parasites. Currently PCR diagnostics are used to distinguish between sibling species. PCR based methods are, however, expensive, time-consuming and their development requires *a priori* DNA sequence information. Here, we evaluated an inexpensive molecular proteomics approach for *Anopheles* species: matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). MALDI-TOF MS is a well developed protein profiling tool for the identification of microorganisms but so far has received little attention as a diagnostic tool in entomology. We measured MS spectra from specimens of 32 laboratory colonies and 2 field populations representing 12 *Anopheles* species including the *A. gambiae* species complex. An important step in the study was the advancement and implementation of a bioinformatics approach improving the resolution over previously applied cluster analysis. Borrowing tools for linear discriminant analysis from genomics, MALDI-TOF MS accurately identified taxonomically closely related mosquito species, including the separation between the M and S molecular forms of *A. gambiae* sensu stricto. The approach also classifies specimens from different laboratory colonies; hence proving also very promising for its use in colony authentication as part of quality assurance in laboratory studies. While being exceptionally accurate and robust, MALDI-TOF MS has several advantages over other typing methods, including simple sample preparation and short processing time. As the method does not require DNA sequence information, data can also be reviewed at any later stage for diagnostic or functional patterns without the need for re-designing and re-processing biological material.

**Citation:** Müller P, Pflüger V, Wittwer M, Ziegler D, Chandre F, et al. (2013) Identification of Cryptic *Anopheles* Mosquito Species by Molecular Protein Profiling. PLoS ONE 8(2): e57486. doi:10.1371/journal.pone.0057486

**Editor:** Luciano A. Moreira, Centro de Pesquisas René Rachou, Brazil

**Received:** October 30, 2012; **Accepted:** January 21, 2013; **Published:** February 28, 2013

**Copyright:** © 2013 Müller et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors have no support or funding to report.

**Competing Interests:** Valentin Pflüger and Dominik Ziegler are employees of Mabritec AG. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials. There are no patents or products in development to declare.

\* E-mail: pie.mueller@unibas.ch

## Introduction

Human malaria is exclusively transmitted by *Anopheles* spp. mosquitoes. Nearly all key malaria mosquito vectors – and many other mosquito species – are members of complexes or groups comprising morphologically indistinguishable sibling species [1]. Exact identification is, therefore, fundamental for understanding vector biology; and hence malaria risk factors and epidemiology. Equally, the success of vector control interventions profoundly relies on accurate information on mosquito populations to establish baseline data for the optimal choice of available tools and to monitor their effectiveness.

A widely discussed case is the *Anopheles gambiae* Giles 1902 species complex comprising at least seven morphologically identical sibling species across Africa [2]. Two of the members of this complex, *A. gambiae* sensu stricto (s.s.) and *A. arabiensis*, are major malaria vectors in sub-Saharan Africa and are found in sympatry over a large geographical range [3]. For malaria

epidemiology and vector control an important aspect is the fact that these species differ in their biology. For example, *A. arabiensis* is more likely to rest outdoors for blood digestion making it a lesser target for indoor-residual spraying (IRS) with insecticides [4,5]. Correct identification has even further-reaching practical consequences where a non-vector species is mistaken for a vector species and vice versa. In some areas, for instance, *A. arabiensis* is sympatric with *A. quadrimaculatus* which – due to its strong preference to cattle [6,7] – is generally considered an unimportant vector species. Moreover, sympatric mosquito species may show different levels of susceptibility to available insecticides for mosquito control (e.g. [8,9]), thus underlining the importance of correct taxonomic classification. Even within *A. gambiae* s.s. further subdivisions are made. These subdivisions were initially defined by karyotypes and called “chromosomal forms” (i.e. Mopti, Bamako, Savanna, Bissau and Forest) [10]. The chromosomal forms can be further grouped into two genetically differentiated “molecular

forms”, M and S, which were originally defined by variations in the intergenic spacer (IGS) and internal transcribed spacer (ITS) ribosomal DNA (rDNA) regions [11] and later found to be separated by other genetic associations [12,13]. The M and S forms have been found to display different ecological tolerances and behaviours adding evidence of reproductive isolation between them [14–18].

The current “gold standard” to distinguish closely related and morphologically indistinguishable specimens is PCR diagnostics. Common to *Anopheles* mosquitoes are sequence variations within the second internal transcribed spacer (ITS2) of the rDNA providing valuable markers for diagnostic assays [1]. PCR protocols such as those developed to distinguish members of the

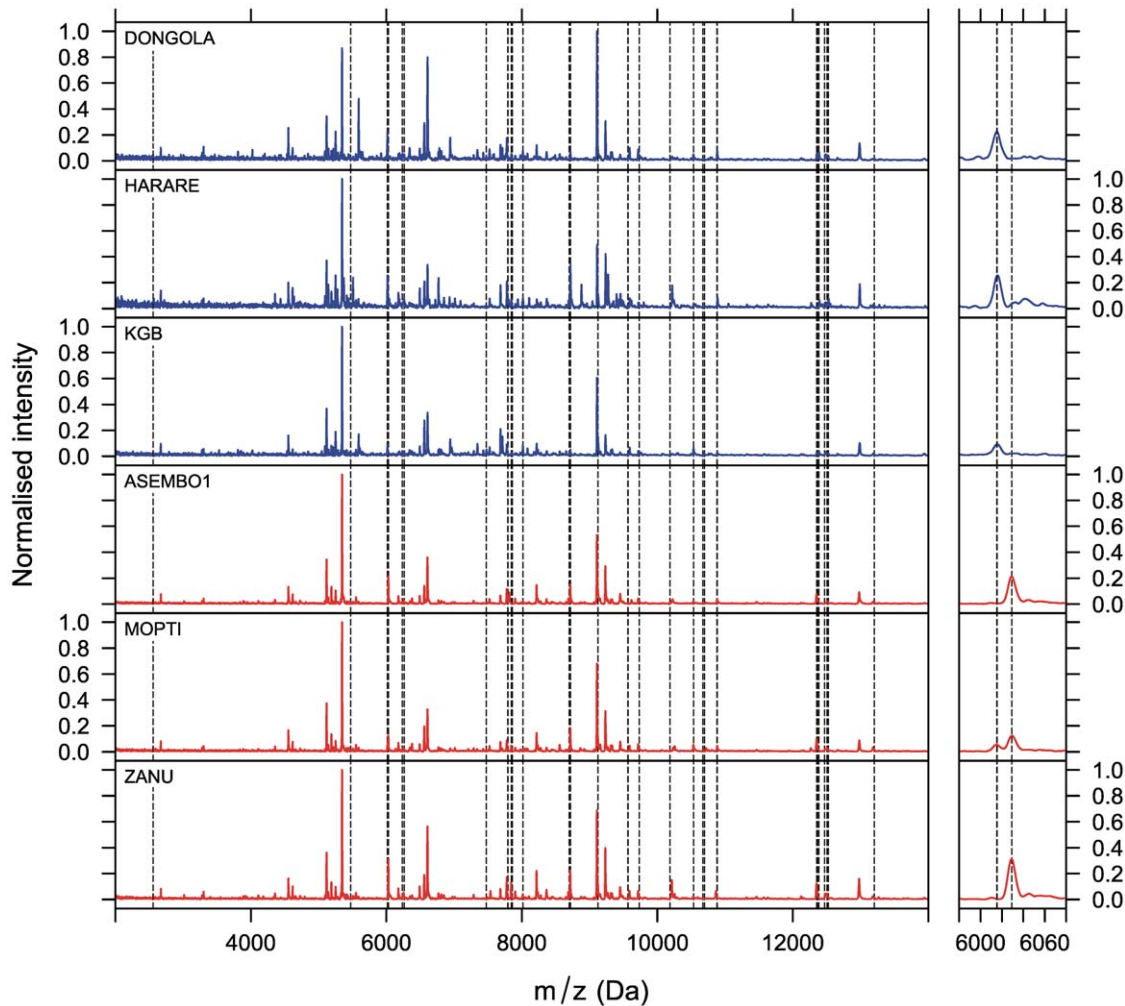
*A. gambiae* complex (e.g. [19–23]) are important tools in basic and applied research. For routine screening they are, however, costly, time-consuming and labour intensive. Species specific PCR is also limited in its flexibility because primers target specific sequences and other potentially important markers may easily be overlooked.

To overcome the drawbacks of classic PCR-based methods, alternative methods have been explored more recently, notably loop-mediated isothermal amplification (LAMP) technique [24] and near-infrared spectroscopy (NIRS) [25,26]. The LAMP technique is a DNA amplification process at a constant temperature using strand displacement reaction, allowing for amplification and detection of a gene in a single step. In contrast, NIRS collects a density distribution of the near-infra-red energy

**Table 1.** Laboratory colonies included in the MALDI-TOF MS analysis.

Species	Molecular form	Colony	Origin	Source <sup>1</sup>
Members of the <i>A. gambiae</i> species complex				
<i>A. quadriannulatus</i> s.s.		SKUQUA	South Africa	MR4 (MRA-761)
		SANGWE	South Africa	MR4
<i>A. merus</i>		OPHANSI	South Africa	MR4 (MRA-803)
		MAF	South Africa	MR4
<i>A. gambiae</i> s.s.	M	MALI-NIH	Mali	MR4 (MRA-860)
	M	MOPTI	Mali	MR4 (MRA-763)
	M	VKS	Burkina Faso	IRD
	M/S mix	VKPER	Benin	IRD
	M/S mix	RSP	Kenya	MR4 (MRA-334)
	S	G3	Gambia	MR4 (MRA-112)
	S	KISUMU1	Kenya	MR4 (MRA-762)
	S	PIMPERENA	Mali	MR4 (MRA-861)
	S	RSP-ST	Kenya	MR4 (MRA-698)
	S	SOUMOUSSO	Burkina Faso	IRD
	S	ZANU	Zanzibar	MR4 (MRA-594)
	S	IN22C+	Isolated from G3	MR4 (MRA-115)
	S (M/S hybrids in males)	ASEMBO1	Kenya	MR4 (MRA-186)
<i>A. arabiensis</i>		BOBO	Burkina Faso	IRD
		DONGOLA	Sudan	MR4 (MRA-856)
		HARARE	Mozambique	MR4
		KGB	Zimbabwe	MR4 (MRA-339)
		SENN	Sudan	MR4 (MRA-764)
Other <i>Anopheles</i> species				
<i>A. stephensi</i>		STE2	India	MR4 (MRA-128)
		STI	India	Swiss TPH
<i>A. quadrimaculatus</i>		ORLANDO	USA	MR4 (MRA-139)
<i>A. minimus</i>		MINIMUS1	Thailand	MR4 (MRA-729)
<i>A. freeborni</i>		F1	USA	MR4 (MRA-130)
<i>A. farauti</i>		FAR1	Papua New Guinea	MR4 (MRA-489)
<i>A. dirus</i>		WRAIR2	Thailand	MR4 (MRA-700)
<i>A. atroparvus</i>		EBRO	Spain	MR4 (MRA-493)
<i>A. albimanus</i>		STECLA	El Salvador	MR4 (MRA-126)
Outgroup				
<i>Aedes aegypti</i>		ROCK	North America	Swiss TPH

<sup>1</sup>MR4: Malaria Research and Reference Reagent Center, VA, USA. Numbers in brackets are MR4 reference numbers; IRD: Institut de Recherche pour le Développement, Montpellier, France; Swiss TPH: Swiss Tropical and Public Health Institute, Basel, Switzerland. Numbers in brackets indicate the MR4 catalogue number. doi:10.1371/journal.pone.0057486.t001



**Figure 1. Examples of MALDI-TOF MS spectra for *Anopheles gambiae sensu stricto* and *A. arabiensis*.** Examples of representative MALDI-TOF MS spectra measured from 3 *A. arabiensis* (blue) and 3 *A. gambiae* s.s. (red) colonies. The spectra were taken from crude suspensions of heads and thoraces in SA solution. The vertical, dashed lines indicate peaks that are characteristic (but not exclusive) for one or the other species. The left panels show the whole spectra between 2 and 14 kDa, while the right panels zoom into two peaks. The two peaks are separated by only a few Daltons. While the left peak is more common in *A. arabiensis*, the right peak is more common in *A. gambiae* s.s. In this representation the peak intensities were normalised against the highest intensity measured in each spectrum.  
doi:10.1371/journal.pone.0057486.g001

absorbed by a sample, which is then explored for characteristics that distinguish biological samples. Both methods may discriminate between the two closely related *A. gambiae* sibling species, *A. gambiae* s.s. and *A. arabiensis* and are potentially valuable methods in areas where the distinction between the two is sufficient [24–26].

A less explored avenue is the use of proteomic data as a taxonomic tool for insects. In contrast, characterisation and identification of microorganisms (*i.e.* bacteria and fungi) using whole cell matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is well established [27]. MALDI-TOF MS produces a mass spectrum that can be compared with reference spectra for rapid species identification and may be even more discriminating than rDNA sequence analysis [28]. Whole cell MALDI-TOF MS requires minimal sample preparation, has very low cost for consumables and produces results within minutes making it ideal for high throughput screening. Through improved hardware and advances in data storage for reference spectra and software solutions, MS

has become a routine approach in the identification of prokaryotes [29,30].

The use of MALDI-TOF MS for the discrimination of arthropod species has been evaluated for the first time almost a decade ago in fruit flies [31] and in aphids [32]. Since then the technique has received rather little attention among entomologists with a few recent exceptions. These include the study of Feltens *et al.* [33] in *Drosophila melanogaster*, by far the most in depth exploration of the method, Kaufmann *et al.* [34,35] in biting midges and Karger *et al.* [36] in ticks.

The prospect of MALDI-TOF MS to discriminate arthropod species motivated us to set out and further explore this technique – together with a computational approach developed for “omics” data – for its use in malaria vector biology. Here, we measured MALDI-TOF MS spectra from 32 *Anopheles* laboratory colonies and 2 field populations representing 12 *Anopheles* species. The *Anopheles* colonies included key members of the *A. gambiae* species complex (*i.e.* *A. gambiae* s.s. of the M and S molecular form, *A. arabiensis*, the zoophagic *A. quadrimaculatus* and one saltwater

**Table 2.** Laboratory colonies included to build the shrinkage discriminant analysis (SDA) models.

Colony	Species	Molecular form	Model 1 <sup>1</sup>	Model 2 <sup>2</sup>	Model 3 <sup>3</sup>
SANGWE	<i>A. quadriannulatus</i> s.s.	-	x		
SKUQUA		-	x		
MAF	<i>A. merus</i>	-	x		
OPHANSI		-	x		
ASEMBO1	<i>A. gambiae</i> s.s.	S	x	x	
G3		S	x	x	
IN22C+		S	x	x	
KISUMU1		S	x	x	
MALI-NIH		M	x	x	
MOPTI		M	x	x	
PIMPERENA		S	x	x	
RSP		M/S mix	x		
RSP-ST		S	x	x	
SOUMOUSO		S	x	x	
VK5		M	x	x	
VKPER		M/S mix	x		
ZANU		S	x	x	
BOBO	<i>A. arabiensis</i>	-	x		x
DONGOLA		-	x		x
HARARE		-	x		x
KGB		-	x		x
SENN		-	x		x

<sup>1</sup>Model 1: SDA model to discriminate between members of the *A. gambiae* species complex.

<sup>2</sup>Model 2: SDA model to discriminate between M and S molecular forms within *A. gambiae* s.s.

<sup>3</sup>Model 3: SDA model to classify specimens to their colony of a single species, i.e. *A. arabiensis*.

doi:10.1371/journal.pone.0057486.t002

species, *A. merus*). This first account of successful implementation of MALDI-TOF MS to identify mosquito vectors of human disease will hopefully prove useful in the field and pave the way for many more related applications in vector biology and entomology.

## Materials and Methods

### Mosquitoes

All mosquito specimens used in the present study were female imagines (adults) with an age of at least two days post eclosure from the pupa. The specimens were either obtained from laboratory colonies or were collected in the field. Specimens from laboratory colonies (Table 1) were either sampled from our own colonies at the Swiss Tropical and Public Health Institute (Swiss TPH) and the Institut de Recherche pour le Développement (IRD; Montpellier and Burkina Faso) or obtained from the Malaria Research and Reference Reagent Center (MR4), VA USA. Field collections were carried out in West Africa in Ladjji, Benin (6°21'10"N, 2°24'30"E) and Soumouso, Burkina Faso (11°01'46"N, 4°02'45"W, see [37]). In Benin, mosquitoes were sampled as larvae in the field and raised to imagines in the laboratory, whereas in Burkina Faso specimens were collected as resting females using aspirators inside human dwellings during the rainy season, in June 2010. In both cases, individual mosquitoes were morphologically identified as members of the *A. gambiae* species complex [38] and shipped to Swiss TPH in 70% ethanol. As found in biting midges [34] our preliminary tests showed that mosquitoes gave sufficient mass spectrometry (MS) signals even if

kept for several months in ethanol (data not shown). Only the heads and thoraces were subjected to MS measurements as abdomens potentially introduce strong bias due to interference by remaining blood meals or changes in physiological status which might interfere with the overall signal – although we cannot fully exclude also some interference with e.g. *Plasmodium* parasites in the salivary glands. The abdomens were kept for molecular typing as described below.

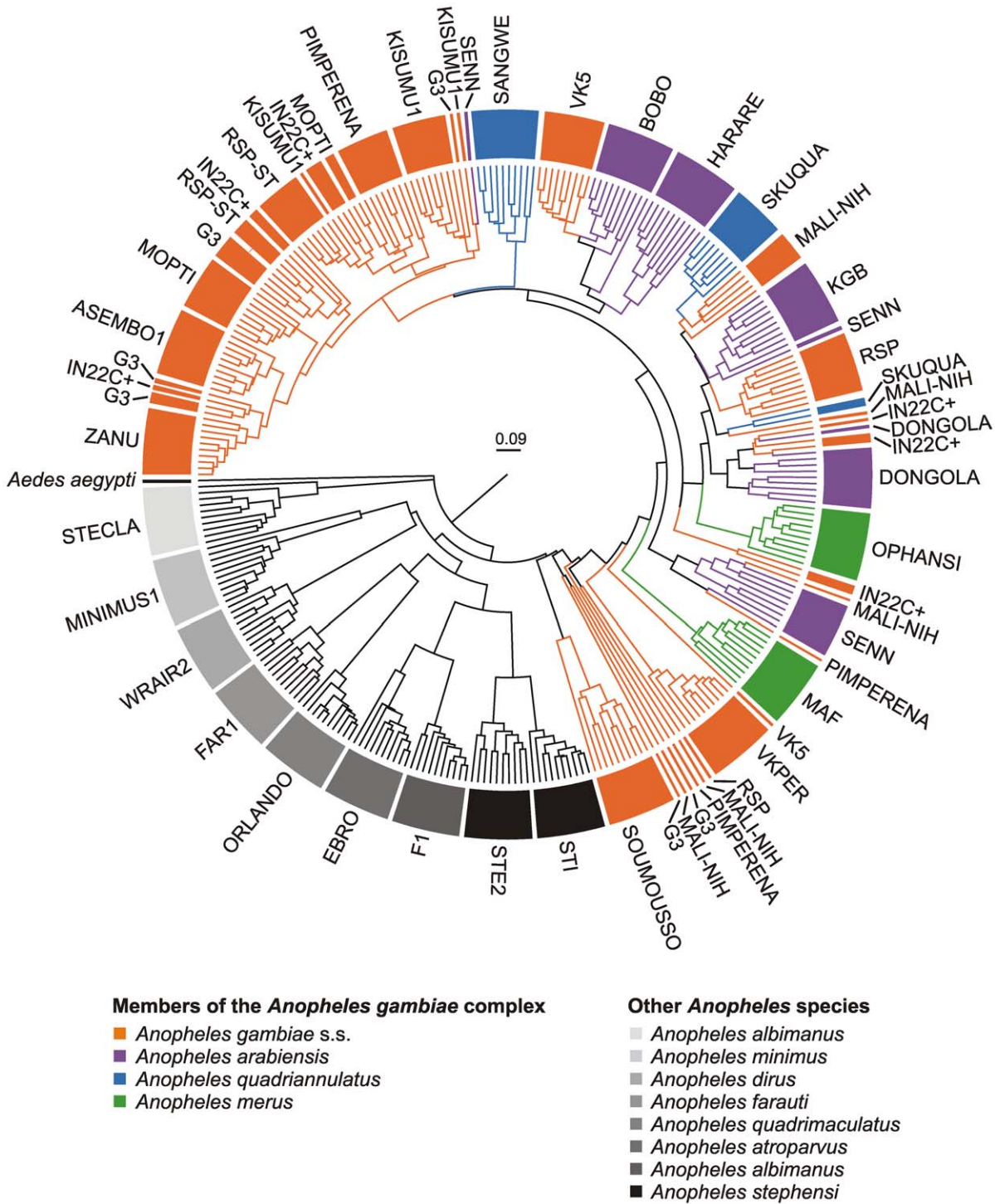
### Molecular Typing

While the MR4 material was regarded as *bona fide*, IRD and Swiss TPH *A. gambiae* laboratory stock and field-caught specimens were genotyped according to the protocol of Wilkins *et al.* [20] with additional primers for simultaneous species and rDNA typing [39] using 1 µl of DNA extracted from the isolated abdomens. DNA was extracted from manually ground (plastic pestles) abdomens using the DNeasy® Blood & Tissue kit (Qiagen, Switzerland) according to the manufacturer's protocol and eluted and stored in 200 µl Buffer AE. PCR products were loaded and run on a 2% agarose gel and visualised by ethidium bromide staining.

### MS Measurements and Data Pre-processing

For the MALDI-TOF MS, dissected head and thoraces were manually ground in a 1.5 ml Eppendorf tube containing 20 µl formic acid (10%). Five µl of the homogenate were then transferred into a new tube containing 7.5 µl saturated sinapic acid (SA) solution. SA solution consisted of 60% acetonitrile and

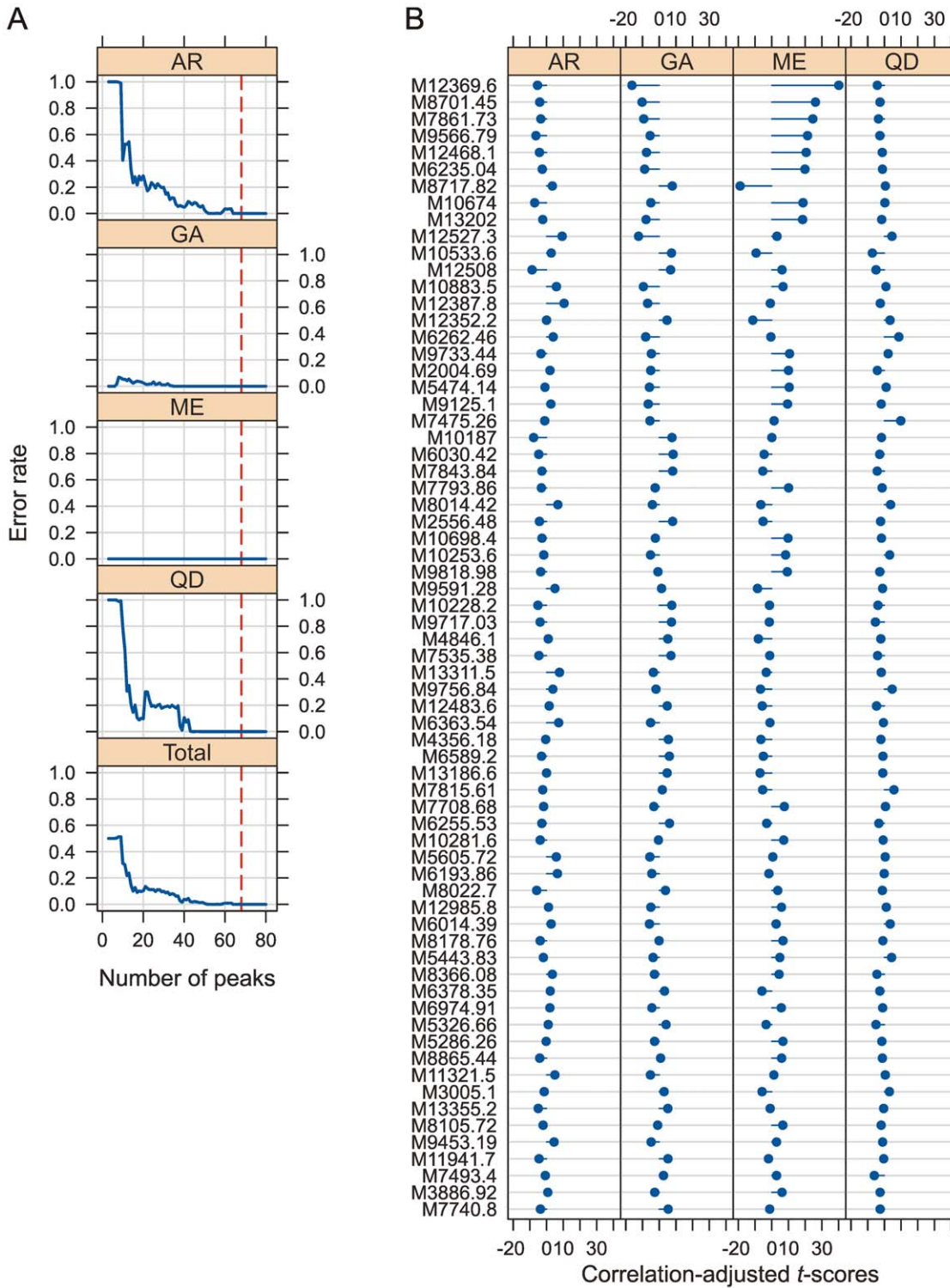




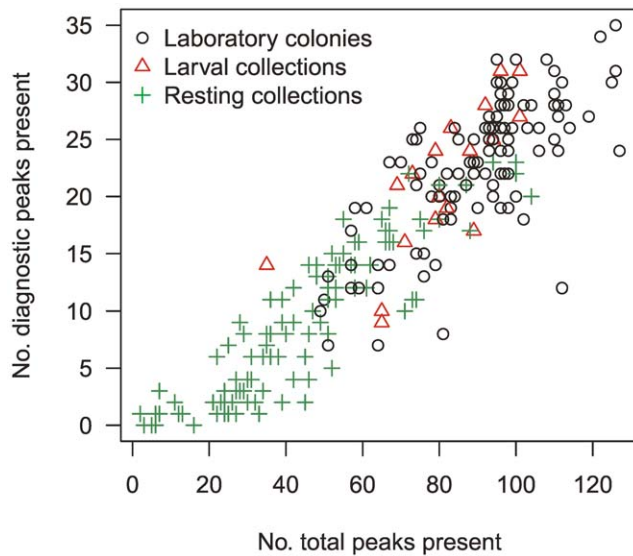
**Figure 2. Dendrogram of hierarchical, unsupervised clustering of binary peaks (presence/absence).** While the *Anopheles* species (complexes) are well separated by the cluster algorithm, the sibling species of the *A. gambiae* complex (coloured lines) do not segregate into well defined clusters. Specimens, both from the same species and colony, are split into different groups. The external branches represent each measured specimen. For each colony spectra from 10 specimens were recorded and included in the cluster analysis. The labels give the names of the colonies (Table 1). The length of the branches corresponds to the size of the Dice similarity coefficient. doi:10.1371/journal.pone.0057486.g002

0.3% trifluoroacetic acid (Sigma-Aldrich, Switzerland). From the matrix suspension 4×1 µl were spotted on a custom made, 48 position steel target plate (Industrietechnik MAB AG, Basel, Switzerland) and air-dried.

Raw spectra were acquired with an Axima™ Confidence MALDI-TOF mass spectrometer (Shimadzu-Biotech Corp., Kyoto, Japan) in the linear, positive mode over a *m/z* range of 2–30 kDa for a total of 1,000 laser shots per spotted sample. The



**Figure 3. Model selection and cross-validation to discriminate between species of the *Anopheles gambiae* complex (Model 1).** (A) The graph shows the error rate from the cross-validation plotted as a function of the number of the ranked peaks included in the SDA model that discriminates between members of the *A. gambiae* species complex. The peaks were ranked (left to right) according to the correlation-adjusted *t*-scores (CAT scores). The vertical, red line shows the 68 peaks chosen for the SDA model. (B) List with the 68 ranked peaks (top equals highest rank) their corresponding CAT scores. The length and direction of the horizontal blue bars represents the CAT scores of the centroid versus the pooled mean and show the influence of a particular peak in differentiating between the groups (Table S2). For example the top peak, M12369.6 has a strong influence in separating *A. merus* from all the other species, emphasised by the length of the bar and the opposite direction from the bars of the other species. In contrast, the tenth peak, M12527.3 has a stronger influence in separating *A. gambiae* s.s. from *A. arabiensis*. AR: *A. arabiensis*; GA: *A. gambiae* s.s.; ME: *A. merus*; QD: *A. quadriannulatus*. doi:10.1371/journal.pone.0057486.g003



**Figure 4. Total number of peaks versus number of diagnostic peaks present in average spectra from the *Anopheles gambiae* species complex.** The number of diagnostic peaks present is associated with the number of total peaks present in an average peak list. The diagnostic peaks refer to the 68 selected peaks to distinguish within the *A. gambiae* species complex (Model 1). The plot suggests that the field specimens collected by aspiration (green crosses) were generally of lower quality (*i.e.* showing fewer peaks) than the specimens that were raised from the larvae; regardless whether from laboratory (black circles) or field caught larvae (red triangles). doi:10.1371/journal.pone.0057486.g004

machine's parameters were chosen by setting the ion source at 20 kV and the extraction delay time at 200 ns. The spectra obtained by the spectrometer were then loaded into Launchpad<sup>TM</sup> 2.8 software (Shimadzu-Biotech Corp., Kyoto, Japan) to create a peak list for each sample spotted on the plate (*i.e.* four lists for each mosquito specimen). The peak lists with size ( $m/z$  value) and intensity values (arbitrary units) were then saved as ASCII text files for further data analysis. The peak lists and additional information are provided in Dataset S1 and Table S1. The software parameters were set to the following values: parent peak cleanup = "advanced scenario"; peak width = "80 chans"; smoothing filter width = "50 chans"; baseline filter width = "500 chans"; peak detection method = "threshold apex". For the "threshold apex" peak detection, the threshold type was set as dynamic and the threshold offset to 0.02 mV with a response factor of 1.2. For external calibration purposes the software also used the spectra of the *Escherichia coli* DH5 alpha strain which was spotted alongside the mosquito samples on each 48-well plate. In addition, internal reference peaks that appear to be highly conserved in mosquitoes (*i.e.*  $m/z$  2670.5, 4554.4, 5115.2, 5217.4, 5328.7, 5345.6, 5371.2, 5551.8, 5591.3, 6560.1, 7683.7, 8560.0, 9109.6, 9234.2, 9453.3, 10255.6 and 16724.8, see also results Figure 1) were used for internal calibration within a range of  $\pm 700$  ppm.

### Unsupervised Cluster Analysis

For the unsupervised cluster analysis, the peak lists, previously saved as ASCII text files, were loaded into SARAMIS<sup>TM</sup> 3.4.1.11 (AnagnosTec, Germany) to create a binary presence/absence table for each peak in the whole data set with columns for peaks and rows for spectra. Peaks were binned within  $\pm 800$  ppm. A custom-written macro in Excel 2002 (Microsoft) then merged the

peak lists into a single "average" peak list for each mosquito specimen. Here, a peak was deemed present if it was detected in at least three out of four lists. The consolidated table was then loaded into the freely available PAST 2.14 software and a dendrogram generated using the "Dice" multivariate clustering algorithm for paired groups [40]. The computed output was exported in nexus file format and displayed with the freely available FigTree 1.3.1 programme (available from <http://tree.bio.ed.ac.uk>) and labels adjusted for visibility with CoreIDRAW 12 (Corel Corporation, 2003).

### Supervised Linear Discriminant Analysis

In addition to the unsupervised clustering approach described above we have also taken a supervised approach using linear discriminant analysis (LDA), an extension to the approach applied to bacteria by Wittwer et al. [41], to find combinations of features (*i.e.* peaks) that separate the "taxonomic" groups at three levels:

Model 1: Classify between members of the *A. gambiae* species complex;

Model 2: Classify *A. gambiae* s.s. into M and S molecular forms; and

Model 3: Classify laboratory colonies of the same *Anopheles* species.

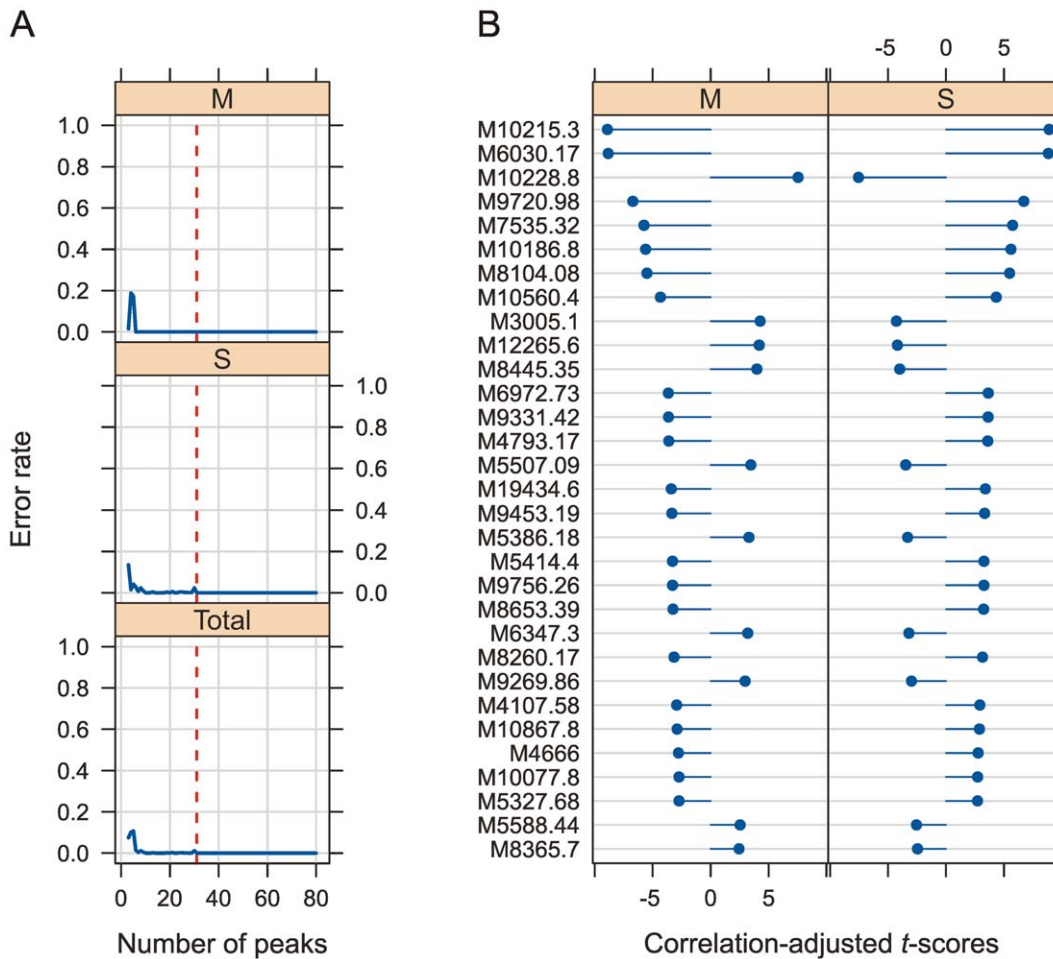
For constructing the LDA model we have chosen the shrinkage discriminant analysis (SDA) procedure described in Ahdesmäki and Strimmer [42] because it addresses well the "small  $n$ , large  $p$ " (*i.e.* few data points, many features) issue and correlation between peaks, a typical feature of MS spectra. Here, training of the classifier is based on James–Stein shrinkage estimates of correlations and variances, where regularisation parameters are chosen analytically without re-sampling and therefore computationally non-intensive. The approach applies a pooled centroids formulation of the multiclass LDA predictor function, in which the relative weights of Mahalanobis-transformed predictors are given by correlation-adjusted  $t$ -scores (CAT scores). The CAT scores allow for simple ranking and selection of peaks.

For the supervised LDA, ASCII files containing the peak list for each measurement (*i.e.* four files per specimen) were imported into the open source statistical software package R version 2.14.1 [43]. The peaks from the original peak lists were dynamically binned with the R package "caMaClass" [44], with a variable bin size between 800–1600 ppm. Here, dynamical binning means that all spectra (*i.e.* peak lists) were aligned to account for small offsets between peak maxima that would still represent the same peak. Peak intensities were  $\log_{10}$ -transformed and then – in line with the unsupervised clustering approach – averaged across the four spectra from the same individual mosquitoes if a signal was present in at least three out of the four spectra. The functions in the R package "sda" [45] were then applied for the SDA as described in Ahdesmäki and Strimmer [42].

For all three models, only laboratory colonies relevant to the specific taxonomic problem were included (Table 2). To build and to test the predictive performance of the SDA model the data was randomly split into two sets so that spectra from five specimens per colony served as the training and validation set for furnishing the SDA classification model and the spectra from the other five mosquitoes per colony formed the true test set for estimating the generalised classification error of the model.

Our approach was to minimise the number of the selected peaks in an attempt to avoid overfitting. The minimum number of peaks was iteratively determined by adding additional peaks with decreasing CAT scores in a stepwise manner to the model, while estimating the error rate from the cross-validation of each set of predictors. The error rate was calculated as the number of failed





**Figure 5. Model selection and cross-validation to distinguish molecular M and S forms in *Anopheles gambiae sensu stricto* (Model 2).** (A) Error rate of the cross-validation plotted as a function of the number of the ranked peaks included in the SDA Model 2 that classifies M and S molecular forms among *A. gambiae* s.s. The peaks were ranked according to the correlation-adjusted t-scores (CAT scores). The vertical, red line shows 31 chosen peaks for the SDA model. (B) The 31 peaks listed on the left were selected on the basis of the smallest number of peaks still providing the lowest error rate shown in (A) and were ranked according to their CAT scores (Table S3). The length and direction of the horizontal blue bars represents the CAT scores of the centroid versus the pooled mean and shows the influence of a particular peak in differentiating between the two molecular forms. M: *A. gambiae* s.s. molecular M form; S: *A. gambiae* s.s. molecular S form. doi:10.1371/journal.pone.0057486.g005

classifications over the total number of successful plus failed classifications committed over 1,000 iterations whereby in every iteration step, a single record from each group (e.g. one from each *A. gambiae* sibling species) of the complete training set was removed by chance and the rest used to fit the classification model.

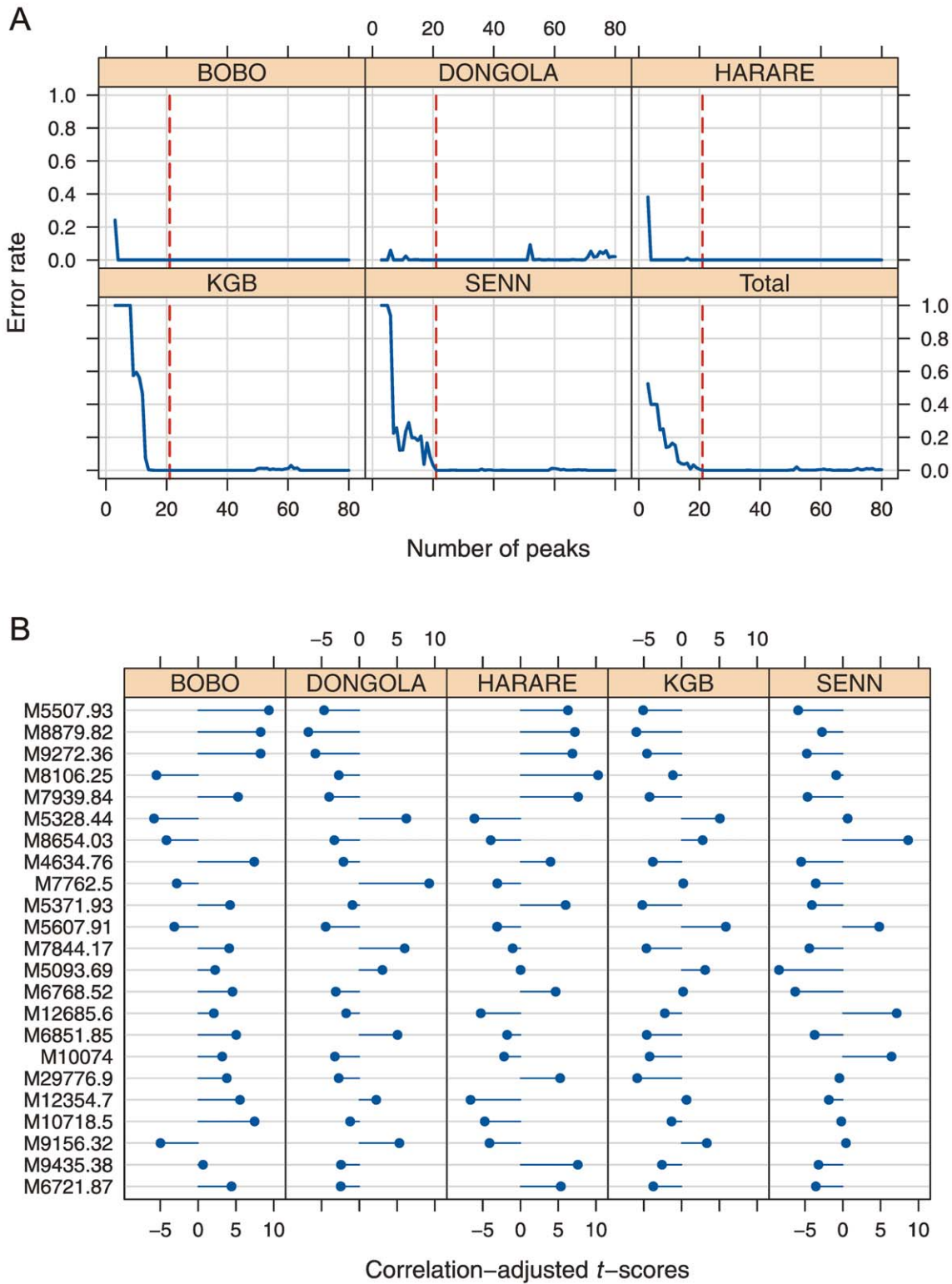
The final classification model was eventually challenged by its application to the specimens that were not used to build and cross-validate the model in order to measure the generalised classification error rate.

**Results and Discussion**

In total, we recorded spectra from 320 laboratory specimens, including 32 *Anopheles* colonies and one *Aedes aegypti* colony (Table 1), and spectra from 125 field-caught specimens that included a mixture of sibling species and molecular forms from two field populations, “Soumouso” in Burkina Faso (20 *A. arabiensis*, 35 *A. gambiae* s.s. M form and 51 *A. gambiae* s.s. S form) and “Ladji” in Benin (19 *A. gambiae* s.s. M form). MALDI-TOF MS spectra measured from heads and thoraces suspended in SA matrix solution produced peaks in the raw spectra with *m/z* values

ranging between 2 and 29.8 kDa with the majority (95% of all peaks) lying within 2 and 15.7 kDa. On average, 121 peaks were detected in a single raw spectrum, ranging between 48 and 187 peaks. This is in the range of MS spectra previously obtained from biting midges using a similar approach [34]. Given that the *A. gambiae* proteome encompasses approximately 13,000 proteins [46] the number of peaks acquired with the current method suggests that only a very limited fraction of the whole proteome is represented. Despite the weak representation of the proteome some peaks appeared to be more characteristic – but not exclusive markers - for certain species even by visual inspection of the unprocessed MS spectra (Figure 1). Intriguingly, in some instances the peaks were separated by only a few Daltons, perhaps reflecting single amino acid substitutions or minor post-transcriptional modifications. Indeed, such small differences were previously found in orthologous neuropeptides from different *Drosophila* species by MALDI-TOF MS [47].





**Figure 6. Model selection and cross-validation for colony authentication in *Anopheles arabiensis* (Model 3).** (A) Error rate as a function of the number of peaks included in the SDA model for five *A. arabiensis* colonies and the total error rate over all colonies. The peaks were ranked according to the correlation-adjusted  $t$ -scores (CAT scores). The vertical, red line shows the 23 peaks chosen for the SDA model (Table S4). (B) Top 23 peaks included in SDA model after they were ranked according to CAT scores (i.e. peak with highest CAT score appears at the top of the list). The length and direction of the horizontal blue bars represents the CAT scores of the centroid versus the pooled mean and shows the influence of a particular peak in differentiating between the colonies.  
doi:10.1371/journal.pone.0057486.g006

## Unsupervised Cluster Analysis

Previous studies that evaluated the feasibility of using whole cell MALDI-TOF MS to distinguish between arthropod species either identified distinct patterns, with similarity within and differences between species, by visual examination [31,32] or hierarchical cluster analysis [33,34,36].

Initially we also set out to use a cluster analysis approach. At first inspection, colonies from the same species (complex) that were reared in different laboratories over many years clustered well together into the same super cluster (Figure 2). A good example is *A. stephensi*. The individual specimens from the two colonies included in this study segregate into two clusters and yet aggregate into one single cluster for that species at the next higher level. This is in contrast to the *A. gambiae* species complex where hierarchical clustering failed to segregate the (four analysed) sibling species within the complex (Figure 2). Conceivably, the intermixture within the species complex mirrors the close relationship among the *A. gambiae* sibling species. In line with the lack of distinct hierarchical clusters there were no unique peaks that would serve as single biomarkers to separate the sibling species.

## Sibling Species Classification within the *Anopheles gambiae* Complex (Model 1)

In an attempt to overcome the poor performance of the unsupervised cluster analysis in discriminating between the *A. gambiae* sibling species, a SDA classification model was evaluated as an alternative. The model (Model 1) was trained using 110 specimens, 5 individual mosquitoes from each of 22 laboratory colonies including 5 *A. arabiensis*, 13 *A. gambiae* s.s., 2 *A. merus* and 2 *A. quadriannulatus* colonies (Table 2). When ranked by the CAT scores, including the top 68 peaks gave a model with zero remaining total error rate in the cross-validation (Figure 3 and Table S2). For estimating the generalised classification error of the final model the other 110 specimens, not used for model building and cross-validation, from the laboratory colonies plus an additional set of 125 field-caught female mosquitoes were classified using Model 1.

Model 1 correctly classified 105 out of 110 laboratory specimens (95% accuracy), while the model's performance for the field-caught mosquitoes was lower with 105 out of 125 specimens correctly identified (84% accuracy). Taking a closer look at the field specimens it turns out that the 19 specimens reared from field caught larvae were all accurately classified, while only 86 out of 106 specimens (*i.e.* 81%) caught by aspiration were correctly identified. Altogether this means that pooling the specimens that were processed in the same way (*i.e.* raised from larvae) show an astonishing accuracy of 96% (124 out of 129). PCR and MALDI-TOF MS scores for each specimen included in the analysis are provided in Table S1.

There might be many – biological and technical – reasons why the specimens from the resting collections performed less well in the MALDI-TOF MS analysis. The specimens – unlike the adults raised from the larval collections – were either blood fed or gravid females or perhaps even carrying pathogens. Therefore, it is expected that a subset of genes would be differentially expressed due to the physiological state of the mosquito (*e.g.* [48–50]). Such differential expression would potentially also affect the number and types of masses detected in the MALDI-TOF spectra. If this was the prime cause of poor performance of Model 1 on the specimens from the resting collections we would predict different masses to come up in the MS spectra. The observation in our data set was, however, a different one. The number of peaks obtained from these specimens was lower than from those raised from larvae. More importantly, the number of diagnostic peaks present

in an average spectrum showed the same relationship for all types of samples (Figure 4). It is, therefore, concluded that a major impact on the poorer performance of Model 1 on these specimens is a quality rather than a biological phenomenon. Indeed, some of the spectra even showed zero peaks (Figure 4). Including the spectra from field specimens in the model did also not improve its performance (data not shown). Perhaps the blood in the abdomens somehow negatively influenced the preservation of the specimens from those resting collections. Although somewhat unfortunate for the current study this is an aspect that can be addressed by optimising and standardising sampling, storage and processing procedures in future studies. A preliminary recommendation would be to separate the abdomens from the head and thoraces prior storage.

## Biological Meaning of the Detected Peaks

An interesting question would be what peptides or proteins the selected peaks actually represent and whether the observed patterns have any biological meaning, but without thorough additional investigation and access to genetic information on the different mosquito taxa the masses themselves only lead to speculation.

Feltens *et al.* [33] investigated some proteins that came up in MALDI-TOF MS profiling in *D. melanogaster* using nano-high-performance liquid chromatography electro spray ionisation tandem MS. Most of them were identified as originating from muscle tissues and mitochondria. As mentioned above, a caveat underlying the MALDI-TOF MS is that the full complement of proteins and peptides cannot be detected. Similarly, it might also miss out on detecting differences between epicuticular lipid profiles that have been described between *A. gambiae* M and S molecular forms as well as *A. arabiensis* as they are below the detection range [51].

## Classification of *Anopheles gambiae* Sensu Stricto M and S Molecular forms (Model 2)

Remarkably, by using the SDA approach it was possible to come up with a SDA model (Model 2) that allows for discriminating between the M and S molecular forms of *A. gambiae* s.s. (Figure 5 and Table S3). Among the 11 laboratory colonies (8 S and 3 M form colonies; Table 2) 50 out of 55 (91%) individuals that were not used to build the model were still correctly identified as either M or S form with 31 peaks in the model (Table S1).

The model failed to discriminate between the M and S molecular forms in the field caught specimens (classification error rate equals 49%). While there were some quality issues as discussed above, additional discrepancies between the rDNA typing method and MALDI-TOF might have arisen due to non-interchangeability between methods as found between several rDNA methods [52,21]. It also appears that the discriminating pattern obtained from the laboratory specimens does not well represent the field caught mosquitoes. This may likely be overcome by adding field specimens into the reference set for building the SDA model.

Although the CAT scores are no direct measurement of phylogenetic distances the computed values qualitatively match the expectation that distances between higher taxa would generally be greater than those of closely related taxa. When comparing the range of CAT scores in Figure 3b (Model 1; discriminating between *A. gambiae* sibling species) to those in Figure 5b (Model 2; discriminating molecular M and S molecular forms) the observed patterns actually meet that prediction.

### Classification of *Anopheles arabiensis* Laboratory Colonies (Model 3)

The SDA Model 3 (Figure 6 and Table S4), classifying specimens of the same species into their colonies of origin, accurately scored 20 out of 25 specimens (80%) among the five *A. arabiensis* laboratory colonies (Table S1). Though 80% accuracy may seem low this is still quite remarkable given that the model is based on only 5 randomly picked individuals per colony. Including more features than the minimum 21 peaks yielded by our inclusion criteria would actually increase accuracy even more. For example, including an additional 7 peaks into the model provides an accuracy of 88% (*i.e.* 22 correctly identified out of 25 individuals). Including more specimens in the training set would also reduce the classification error (data not shown).

### How does the Approach Compare to Other Species Diagnostic Tools?

Assays widely used to distinguish between members of the *A. gambiae* species complex are PCR diagnostics based on sequence variations within the ITS2 of the rDNA [53,19–21]. These assays require several steps in processing mosquito specimens; DNA extraction, PCR amplification and finally visualisation of the amplicon. Altogether these steps are labour and cost intensive. Even just the consumables for a single extraction followed by a PCR step can easily be 100 times more expensive than the consumables for a crude MALDI-TOF measurement - not even mentioning the increased labour costs and processing time.

An obstacle for the use of MALDI-TOF MS, particularly in disease endemic countries could be the large capital outlay required for acquiring and running a spectrometer and the need for specially trained personnel. Once a sound data basis and the analytical tools are implemented into automated systems this technique may, however, become an accessible tool for a wider community and a valuable alternative for large scale screening programmes.

For smaller projects an interesting alternative is LAMP, loop-mediated isothermal amplification technique as it uses little laboratory equipment and is much faster than current PCR methods [24]. The method, however, still requires knowledge of species specific differences to develop the assay and design of sequence specific primers in the case of multiple species is challenging.

MALDI-TOF MS has the great advantage that *a priori* knowledge of sequence variations is not needed and once spectra are acquired data can be revisited at any time *in silico* as already demonstrated here. For example, separating *A. arabiensis* from *A. gambiae* s.s. the same data could be used to first separate between the species and then to further classify *A. gambiae* s.s. into M and S molecular forms. Similarly this would allow for re-running the analysis should taxonomy change in one or the other way.

Another approach that has been proposed to discriminate between *A. gambiae* s.s. and *A. arabiensis* specimens is NIRS, near infrared spectroscopy [25]. Similar to our approach spectra are recorded and explored for discriminant patterns. The spectra themselves are, however, less conducive in drawing conclusion as to what causes them and what makes the differences between classes due to the complex nature of the spectra. An association between observed patterns and specific chemical components is extremely difficult if not impossible.

In summary, combining acquisition of MALDI-TOF MS spectra and statistical analytical tools to classify mosquito specimens appears very promising for *Anopheles* research and routine surveys for vector control programmes and, most importantly,

entomology in general. A strength of the SDA algorithm is also that it takes into account peak intensities adding to the possibilities in discriminating patterns. The taxonomic classes are separated by patterns rather than single diagnostic peaks. It is expected that models including more reference specimens together with better storage and/or processing procedures will increase accuracy and add further value to this technology “repurposed” from microbiology. Furthermore, purification of the protein extracts might yield larger numbers of peaks similar to those found in *D. melanogaster* [33] that would also allow for investigating the nature of the peaks themselves.

### Conclusions

The present study shows that MALDI-TOF MS reliably discriminates between anopheline mosquito species - even at the sub-species level. Present data suggests that even colony-specific patterns are resolved and that the technique may be used beyond simple species typing including stock authentication or perhaps the detection of population structures in field-caught mosquitoes. While being accurate and robust MALDI-TOF MS has several additional advantages over other typing methods, including simple sample preparation, short processing time and low consumable costs - providing results rapidly and economically. The workflow can easily be standardised and automated allowing for cost-effective high throughput mass screening. As the method does not require DNA sequence information about the mosquito, data can be reviewed at any later stage for diagnostic or functional patterns. As only parts of the animal are needed the remaining parts can be subjected to additional analysis on DNA or protein extracts of the same individual. This method has the potential to become an invaluable tool for many applications in vector biology and control including routine species identification, colony authentication, population genetics or even the detection of trait-specific markers including insecticide resistance. These and other possibilities are currently being further explored in our laboratories.

### Supporting Information

**Table S1 MALDI-TOF MS classification results.** The table shows the specimens included in the three different models and whether the models classified the individuals correctly (TRUE) or wrongly (FALSE) against the reference (*i.e.* morphology, PCR score and information provided by MR4). PEAKLIST: file name of the MALDI-TOF MS peak list. LABEL: name of the mosquito colony or the field population. MR4.ID: MR4 catalogue number. ORIGIN: tells whether the specimen was collected in the field or originating from a laboratory colony. SPECIMEN: is the specimen number of a colony (1 through to 10) or field population. SPECIES: mosquito species. MOL.FORM: molecular form in the case of *A. gambiae* s.s. INCLUSION.MODEL.1: indicates whether the spectrum/specimen was included in building Model 1. CLASSIFICATION.MODEL.1: states whether the specimen was classified correctly (TRUE), wrongly (FALSE) or was not included in the classification (NA). INCLUSION.MODEL.2: same as above for Model 2. CLASSIFICATION.MODEL.2: same as above for Model 2. INCLUSION.MODEL.3: same as above for Model 3. CLASSIFICATION.MODEL.3: same as above for Model 3. (CSV)

**Table S2 CAT scores for the SDA model discriminating the *Anopheles gambiae* species complex (Model 1).** The table lists the computed correlation-adjusted *t*-scores (CAT scores) of the mean versus the pooled mean for each predictor variable (*i.e.* peak) and centroid. Score: The sum of the squared CAT scores

across groups which determines the overall ranking of the peaks. CAT.AR, CAT.GA, CAT.ME and CAT.QD are the CAT scores of the centroid versus the pooled mean for each group and peak. LFDR: The local false discovery rate computed for each peak. HC: The higher criticism score computed for each peak. (CSV)

**Table S3 CAT scores for the SDA model discriminating M and S molecular forms in *Anopheles gambiae* s.s. (Model 2).** The table lists the computed correlation-adjusted *t*-scores (CAT scores) of the mean versus the pooled mean for each predictor variable (*i.e.* peak) and centroid. Score: The sum of the squared CAT scores across groups which determines the overall ranking of the peaks. CAT.M and CAT.S are the CAT scores of the centroid versus the pooled mean for each group and peak. LFDR: The local false discovery rate computed for each peak. HC: The higher criticism score computed for each peak. (XLS)

**Table S4 CAT scores for the SDA model discriminating the *Anopheles arabiensis* colonies (Model 3).** The table lists the computed correlation-adjusted *t*-scores (CAT scores) of the mean versus the pooled mean for each predictor variable (*i.e.* peak) and centroid. Score: The sum of the squared CAT scores across groups which determines the overall ranking of the peaks. CAT.BOBO, CAT.DONGOLA, CAT.HARARE, CAT.KGB and CAT.SENN are the CAT scores of the centroid versus the pooled mean for each group and peak. LFDR: The local false discovery rate computed for each peak. HC: The higher criticism score computed for each peak.

## References

- Collins FH, Paskewitz SM (1996) A review of the use of ribosomal DNA (rDNA) to differentiate among cryptic *Anopheles* species. *Insect Mol Biol* 5: 1–9.
- White BJ, Collins FH, Besansky NJ (2011) Evolution of *Anopheles gambiae* in relation to humans and malaria. *Annu Rev Ecol Syst* 42: 111–132.
- Coetzee M, Craig M, le Sueur D (2000) Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. *Parasitol Today (Regul Ed)* 16: 74–77.
- Githeko AK, Service MW, Mbogo CM, Atieli FK, Juma FO (1994) Origin of blood meals in indoor and outdoor resting malaria vectors in western Kenya. *Acta Trop* 58: 307–316.
- Bayoh MN, Mathias DK, Odiere MR, Mutuku FM, Kamau L, et al. (2010) *Anopheles gambiae*: historical population decline associated with regional distribution of insecticide-treated bed nets in western Nyanza Province, Kenya. *Malar J* 9: 62.
- Dekker T, Takken W (1998) Differential responses of mosquito sibling species *Anopheles arabiensis* and *An. quadrimaculatus* to carbon dioxide, a man or a calf. *Med Vet Entomol* 12: 136–140.
- White GB (1974) *Anopheles gambiae* complex and disease transmission in Africa. *Trans R Soc Trop Med Hyg* 68: 278–301.
- Ramphul U, Boase T, Bass C, Okedi LM, Donnelly MJ, et al. (2009) Insecticide resistance and its association with target-site mutations in natural populations of *Anopheles gambiae* from eastern Uganda. *Trans R Soc Trop Med Hyg* 103: 1121–1126.
- Hargreaves K, Kockemoer LL, Brooke BD, Hunt RH, Mthembu J, et al. (2000) *Anopheles funestus* resistant to pyrethroid insecticides in South Africa. *Med Vet Entomol* 14: 181–189.
- della Torre A, Costantini C, Besansky NJ, Caccone A, Petrarca V, et al. (2002) Speciation within *Anopheles gambiae* - the glass is half full. *Science* 298: 115–117.
- della Torre A, Fanello C, Akogbeto M, Dossou-yovo J, Favia G, et al. (2001) Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. *Insect Mol Biol* 10: 9–18.
- White BJ, Cheng C, Simard F, Costantini C, Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Mol Ecol* 19: 925–939.
- Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol* 3: e285.
- Gimonneau G, Pombi M, Dabiré RK, Diabaté A, Morand S, et al. (2012) Behavioural responses of *Anopheles gambiae* sensu stricto M and S molecular form larvae to an aquatic predator in Burkina Faso. *Parasit Vectors* 5: 65.
- Lehmann T, Diabaté A (2008) The molecular forms of *Anopheles gambiae*: a phenotypic perspective. *Infect Genet Evol* 8: 737–746.
- Gimonneau G, Pombi M, Choisy M, Morand S, Dabiré RK, et al. (2012) Larval habitat segregation between the molecular forms of the mosquito *Anopheles gambiae* in a rice field area of Burkina Faso, West Africa. *Med Vet Entomol* 26: 9–17.
- Gimonneau G, Bouyer J, Morand S, Besansky NJ, Diabaté A, et al. (2010) A behavioral mechanism underlying ecological divergence in the malaria mosquito *Anopheles gambiae*. *Behavioral ecology: official journal of the International Society for Behavioral Ecology* 21: 1087–1092.
- Diabaté A, Dao A, Yaro AS, Adamou A, Gonzalez R, et al. (2009) Spatial swarm segregation and reproductive isolation between the molecular forms of *Anopheles gambiae*. *Proc Biol Sci* 276: 4215–4222.
- Fanello C, Santolamazza F, della Torre A (2002) Simultaneous identification of species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP. *Med Vet Entomol* 16: 461–464.
- Wilkins EE, Howell PI, Benedict MQ (2006) IMP PCR primers detect single nucleotide polymorphisms for *Anopheles gambiae* species identification, Mopti and Savanna rDNA types, and resistance to dieldrin in *Anopheles arabiensis*. *Malaria Journal* 5: 125.
- Bass C, Williamson MS, Field LM (2008) Development of a multiplex real-time PCR assay for identification of members of the *Anopheles gambiae* species complex. *Acta Trop* 107: 50–51.
- Scott JA, Brogdon WG, Collins FH (1993) Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *Am J Trop Med Hyg* 49: 520–529.
- Favia G, della Torre A, Bagayoko M, Lanfrancotti A, Sagnon N, et al. (1997) Molecular identification of sympatric chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive isolation. *Insect Mol Biol* 6: 377–383.
- Bonizzoni M, Afrane Y, Yan G (2009) Loop-mediated isothermal amplification (LAMP) for rapid identification of *Anopheles gambiae* and *Anopheles arabiensis* mosquitoes. *Am J Trop Med Hyg* 81: 1030–1034.
- Mayagaya VS, Michel K, Benedict MQ, Killeen GF, Wirtz RA, et al. (2009) Non-destructive determination of age and species of *Anopheles gambiae* s.l. using near-infrared spectroscopy. *Am J Trop Med Hyg* 81: 622–630.
- Sikulu M, Killeen GF, Hugo LE, Ryan PA, Dowell KM, et al. (2010) Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors. *Parasit Vectors* 3: 49.
- Giebel R, Worden C, Rust SM, Kleinheinz GT, Robbins M, et al. (2010) Microbial fingerprinting using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) applications and challenges. *Adv Appl Microbiol* 71: 149–184.
- Dickinson DN, La Duc MT, Satomi M, Winefordner JD, Powell DH, et al. (2004) MALDI-TOF MS compared with other polyphasic taxonomy approaches for the identification and classification of *Bacillus pumilus* spores. *J Microbiol Methods* 58: 1–12.



29. Sauer S, Freiwald A, Maier T, Kube M, Reinhardt R, et al. (2008) Classification and identification of bacteria by mass spectrometry and computational analysis. *PLoS ONE* 3: e2843.
30. Freiwald A, Sauer S (2009) Phylogenetic classification and identification of bacteria by mass spectrometry. *Nat Protoc* 4: 732–742.
31. Campbell PM (2004) Species differentiation of insects and other multicellular organisms using matrix-assisted laser desorption/ionization time of flight mass spectrometry protein profiling. *Systematic Entomology* 30: 186–190.
32. Perera MR, Vargas RDF, Jones MGK (2005) Identification of aphid species using protein profiling and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Entomol Exp Appl* 117: 243–247.
33. Feltens R, Görner R, Kalkhof S, Gröger-Arndt H, Von Bergen M (2010) Discrimination of different species from the genus *Drosophila* by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry. *BMC Evolutionary Biology* 10: 95.
34. Kaufmann C, Ziegler D, Schaffner F, Carpenter S, Pflüger V, et al. (2011) Evaluation of matrix-assisted laser desorption/ionization time of flight mass spectrometry for characterization of *Culicoides nubeculosus* biting midges. *Med Vet Entomol* 25: 32–38.
35. Kaufmann C, Schaffner F, Ziegler D, Pflüger V, Mathis A (2012) Identification of field-caught *Culicoides* biting midges using matrix-assisted laser desorption/ionization time of flight mass spectrometry. *Parasitology* 139: 248–258.
36. Karger A, Kampen H, Bettin B, Dautel H, Ziller M, et al. (2012) Species determination and characterization of developmental stages of ticks by whole-animal matrix-assisted laser desorption/ionization mass spectrometry. *Ticks Tick Borne Dis* 3: 78–89.
37. Badolo A, Traoré A, Jones CM, Sanou A, Flood L, et al. (2012) Three years of insecticide resistance monitoring in *Anopheles gambiae* in Burkina Faso: resistance on the rise? *Malaria journal* 11: 232.
38. Gillies MT, Coetzee M (1987) A supplement to the anophelinae of Africa south of the Sahara (Afrotropical Region). South African Institute for Medical Research. 143 p.
39. Dotson E (2011) MR4: Methods in Anopheles Research. 2nd ed. Atlanta USA: Centres of Disease Control. 343 p.
40. Hammer O, Harper DA, Ryan PD (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeo Electronica* 4: 9.
41. Wittwer M, Heim J, Schär M, Dewarrit G, Schürch N (2011) Tapping the potential of intact cell mass spectrometry with a combined data analytical approach applied to *Yersinia* spp.: detection, differentiation and identification of *Y. pestis*. *Syst Appl Microbiol* 34: 12–19.
42. Ahdesmäki M, Strimmer K (2010) Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Ann Appl Stat* 4: 503–519.
43. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria. p. Available: <http://www.R-project.org>.
44. Tuszynski J (2010) caMassClass: Processing & Classification of Protein Mass Spectra (SELDI) Data. Available: <http://CRAN.R-project.org/package=caMassClass>.
45. Ahdesmäki M, Zuber V, Strimmer K (2011) Shrinkage Discriminant Analysis and CAT Score Variable Selection. Available: <http://CRAN.R-project.org/package=sda>.
46. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, et al. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298:
47. Wegener C, Gorbashov A (2008) Molecular evolution of neuropeptides in the genus *Drosophila*. *Genome Biol* 9: R131.
48. Das S, Radtke A, Choi Y-J, Mendes A, Valenzuela J, et al. (2010) Transcriptomic and functional analysis of the *Anopheles gambiae* salivary gland in relation to blood feeding. *BMC Genomics* 11: 566.
49. Dana A, Hong Y, Kern M, Hillenmeyer M, Harker B, et al. (2005) Gene expression patterns associated with blood-feeding in the malaria mosquito *Anopheles gambiae*. *BMC Genomics* 6: 5.
50. Félix RC, Müller P, Ribeiro V, Ranson H, Silveira H (2010) *Plasmodium* infection alters *Anopheles gambiae* detoxification gene expression. *BMC Genomics* 11:
51. Caputo B, Dani FR, Horne GL, N'Fale S, Diabate A, et al. (2007) Comparative analysis of epicuticular lipid profiles of sympatric and allopatric field populations of *Anopheles gambiae* s.s. molecular forms and *An. arabiensis* from Burkina Faso (West Africa). *Insect Biochem Mol Biol* 37: 389–398.
52. Santolamazza F, Caputo B, Calzetta M, Vicente JL, Mancini E, et al. (2011) Comparative analyses reveal discrepancies among results of commonly used methods for *Anopheles gambiae* molecular form identification. *Malar J* 10: 215.
53. Santolamazza F, Della Torre A, Caccone A (2004) Short report: A new polymerase chain reaction-restriction fragment length polymorphism method to identify *Anopheles arabiensis* from *An. gambiae* and its two molecular forms from degraded DNA templates or museum samples. *Am J Trop Med Hyg* 70: 604–606.