RESEARCH ARTICLE

A gene expression panel for estimating age in males and females of the sleeping sickness vector *Glossina morsitans*

Eric R. Lucas[1]*, **Alistair C. Darby**[2], **Stephen J. Torr**[1], **Martin J. Donnelly**[1,3]

1 Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **2** Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom, **3** Wellcome Sanger Institute, Cambridge, United Kingdom

* eric.lucas@lstmed.ac.uk

OPEN ACCESS

Citation: Lucas ER, Darby AC, Torr SJ, Donnelly MJ (2021) A gene expression panel for estimating age in males and females of the sleeping sickness vector *Glossina morsitans*. PLoS Negl Trop Dis 15(9): e0009797. https://doi.org/10.1371/journal.pntd.0009797

Editor: Felix Hol, Institut Pasteur, FRANCE

Received: June 3, 2021

Accepted: September 8, 2021

Published: September 23, 2021

Data Availability Statement: Analysis scripts, qPCR raw data and RNAseq read counts are available on GitHub (https://github.com/EricRLucas/TsetseAgeMarkers). Raw sequencing have been submitted to ENA shotgun sequencing archive (project accession PRJEB47261).

Funding: This work was supported by a Liverpool School of Tropical Medicine internal grant (Director's Catalyst Fund, https://lstmed.ac.uk) to ERL and a Medical Research Council, UK (MR/T001070/1, https://mrc.ukri.org/) grant to MJD and ERL. ST received support from the UK's

Abstract

Many vector-borne diseases are controlled by methods that kill the insect vectors responsible for disease transmission. Recording the age structure of vector populations provides information on mortality rates and vectorial capacity, and should form part of the detailed monitoring that occurs in the wake of control programmes, yet tools for obtaining estimates of individual age remain limited. We investigate the potential of using markers of gene expression to predict age in tsetse flies, which are the vectors of deadly and economically damaging African trypanosomiases. We use RNAseq to identify candidate expression markers, and test these markers using qPCR in laboratory-reared *Glossina morsitans morsitans* of known age. Measuring the expression of six genes was sufficient to obtain a prediction of age with root mean squared error of less than 8 days, while just two genes were sufficient to classify flies into age categories of ≤15 and >15 days old. Further testing of these markers in field-caught samples and in other species will determine the accuracy of these markers in the field.

Author summary

Many insect-borne diseases are controlled by methods that kill the insects responsible for disease transmission. Estimating the age structure of populations of disease-transmitting insects provides information on mortality rates and on the capacity of the population to maintain disease transmission, and should form part of the detailed monitoring that occurs in the wake of control programmes, yet tools for obtaining estimates of individual age remain limited. We investigate the potential of using markers of gene expression to predict age in tsetse flies, which are the vectors of deadly and economically damaging African trypanosomiases, such as sleeping sickness. We use RNA sequencing to identify candidate genes, and test their accuracy in laboratory-reared tsetse flies of known age. Measuring the expression of six genes was sufficient to obtain a prediction of age with root mean squared error of less than 8 days, while just two genes were sufficient to classify flies into age categories of ≤15 and >15 days old. Further testing of these markers in field-

Biotechnology and Biological Sciences Research Council (grant numbers: BB/S01375X/1, BB/S00243X/1, BB/P005888/1, https://bbsrc.ukri.org/) and the Bill and Melinda Gates Foundation (INV-001785, OPP1155293, https://www.gatesfoundation.org).The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

caught samples and in other species will determine the accuracy of these markers in the field.

1. Introduction

Vector-borne diseases represent major threats to health and livelihood world-wide, being directly responsible for 680,000 deaths annually [1], as well as causing huge economic damage to livestock [2,3]. Control of the vectors that transmit these diseases is an integral tool for reducing disease burden [4]. The metric of success for these control programmes is a reduction in disease burden in the host population. However, when vector control is accompanied by other interventions such as screening and treating the host population for the disease, the contribution of vector control to the subsequent reduction of disease can be hard to determine [5]. Conversely, while the impact on the vector population may not bear a simple relationship to disease burden, it is a direct outcome of vector control. Control efforts should thus be accompanied by detailed monitoring of the targeted vector populations, to estimate impact, to monitor population recovery and to understand the transmission dynamics of the disease. Mostly, monitoring currently relies on counting the number of vectors caught in sentinel traps, which can be greatly affected by trapping method, effort and efficacy, and may only partly reflect the ability of the vector population to transmit disease [6].

One aspect of vector monitoring that has been particularly challenging is the quantification of the age-distribution (demographics) of natural populations [7–9]. Estimating vector age is important for two reasons. First, it can provide a measure of the effectiveness of vector control because increased adult mortality should lead to a younger population age structure. Importantly, this measure of control effectiveness is independent of catch size and trapping effort because only the distribution of age needs to be known. Second, in most cases, the probability that an individual vector is infectious for a given disease increases with age [10,11]. Before transmitting the disease, vectors first need to have taken an infected blood meal, and there is then typically a delay between acquisition of infection and onward transmission due to the need for the pathogen to replicate and/or mature. Age grading is therefore useful to determine the proportion of individuals old enough to transmit disease.

Tsetse flies (genus *Glossina*) are the vectors of Human African Trypanosomiasis (HAT, or sleeping sickness) and Animal African Trypanosomiasis (AAT, or nagana). HAT is, without treatment, a fatal disease endemic to sub-Saharan Africa [12], while AAT presents a major economic burden to rural communities by affecting livestock [2]. Being a disease primarily of animals and with reservoirs across multiple species, AAT cannot be controlled through treatment alone and is thus highly dependent on vector control [13]. HAT can be more readily controlled through treatment of infected humans, but both the anthroponotic “Gambian” HAT and the zoonotic “Rhodesian” HAT also require some measure of tsetse control to reduce transmission [14]. The choice of tsetse control method depends largely on the ecology and feeding habits of the species being targeted, as well as on local practicalities, but most methods rely on the use of insecticides to directly kill the flies, often applied to baited targets or cattle [13], imposing increased mortality that should translate into a shift in age structure.

G. morsitans morsitans is a major vector of AAT in East and Southern Africa and can also transmit HAT [15]. Catch rates of this species in the wake of vector control can be extremely low [16–18], making it particularly challenging to conduct ongoing monitoring of important populations. It is therefore all the more important to extract as much information as possible from the limited number of flies obtained.

As is the case for all insect vectors, a means to accurately determine the age of tsetse flies is a valuable but elusive goal, and current methods have many shortcomings. Laborious ovary dissections can be used to age females up to their fourth ovarian cycle [19], but this technique requires specialist dissection skills and cannot be applied to males, despite males being at least as competent at transmission as females, and perhaps more so [15,20]. Estimates of age based on wing damage [21] or analysis of pteridines have also been used [22,23], but experience in practical applications has shown that measurements in the field vary enormously (for example in mosquitoes [24,25]) and cannot be used to reliably estimate age on an individual basis [26].

Here we explore the value of using gene expression to estimate age in tsetse flies. This method has previously been tested in mosquitoes [7,8], with encouraging results, but has yet to be applied in tsetse. We use laboratory-reared *G. morsitans* as a proof of concept, and show that measuring the expression of just six genes can estimate the age of both male and female tsetse flies with a root mean squared error of less than 8 days. We also trained models to classify tsetse into those younger or older than 15 days, since flies younger than 15 days are unlikely to harbour a mature trypanosome infection [15], and found that just two genes are sufficient for 95% accurate classification.

2. Methods

2.1. Sample collection and RNA extraction

G. morsitans morsitans individuals were collected from colonies maintained at the Liverpool School of Tropical Medicine. Colonies are kept in meshed boxes (cages) at 26˚C ± 2˚C and 72 ± 4% humidity, with a 12hr light-dark photoperiod, and fed three times per week using defibrinated horse blood (TCS Biosciences Ltd., Buckingham, UK) provided through silicon-membrane feeders. Pupae are regularly collected and allowed to emerge to form new cages. Each fly cage contains flies which eclosed over a 2–3 day window, and thus the age of all flies in the cage are known to a precision of either 2 or 3 days. The ages reported here are the middle of the age range (eg: a fly aged 13–15 days or 13–16 days is reported as 14 days old). The age of the samples ranged from 2 to 62 days. While reproductive status of females was not measured precisely, we tried to include a range of physiological states (based on visual inspection of the size of the abdomen) within each age group, so that genes could be identified that are predictive of age in spite of variation caused by the ovarian cycle. Overall, 505 flies were collected (301 female and 204 male, S1 Data).

For sample collection, fly cages were briefly transferred to a cold room (4˚C) where flies to be collected were removed from the cage once quiescent and decapitated. Heads were placed into RNAlater and stored at -20˚C. In case repeated exposure to the cold room created alterations in gene expression, we minimised this exposure by never collecting flies from a given cage more than three times over the course of the experiment. No more than two flies were collected from a cage on a given day, for three reasons. Firstly, we wanted to make sure that flies were obtained from a range of different cages in order to avoid issues of results being confounded by cage of origin (such as an infection specific to one cage of flies). We therefore never obtained more than six flies from a single cage over the course of the experiment. Second, we wanted to minimise the time that samples spent at temperatures above -20˚C after death, limiting the number of samples that could be collected in a single sitting. Third, all flies were collected at the same approximate time of day (morning) to minimise gene expression variation due to circadian cycles [27], limiting the number of collections that could be performed on the same day.

RNA was extracted from individual fly heads. Single heads contain enough material for RNA sequencing and can easily be removed without the need for precise dissection, providing a quick and convenient tissue for sampling. We avoided the abdomen because of the important effect that sex and the ovarian cycle would have on gene expression in these tissues. RNA extractions were performed using PicoPure kits (Arcturus), increasing the volume of extraction buffer and alcohol to 120μl. cDNA libraries were prepared using SuperScript III Reverse Transcriptase (Invitrogen).

2.2. Sequencing

cDNA libraries from 22 male and 28 female individual flies ranging in age from 2 to 62 days post-eclosion (Fig 1, S1 Data) were sent to the Liverpool Centre for Genomic Research (CGR) for 150bp paired-end sequencing on an Illumina HiSeq 4000 sequencer. Strand-specific library preparation was performed using NEBNext poly A selection and Ultra Directional RNA library preparation kits, producing an average of 23.8 million reads per sample. Reads were then trimmed as part of the CGR's genomic pipeline using Cutadapt version 1.2.1 [28] with option -O 3 to remove Illumina adapter sequences, and Sickle version 1.2 (https://github.com/najoshi/sickle/releases/tag/v1.2) with a minimum window quality score of 20. Reads shorter than 20 bp after trimming were removed and subsequently unpaired reads were excluded. Data were quality checked using FastQC [29] before analysis.

2.3. RNAseq analysis

Trimmed reads were aligned to the GmorY1.9 genome using STAR aligner version 2.7.0 [30] using the—quantMode GeneCounts option to obtain mapping counts for each gene.

Differential expression analysis was performed using the R package *EdgeR [31]*, with library size normalisation performed using Trimmed Mean of M-values [32] and dispersion calculated with trended and tag-wise estimates. Genes with fewer than 10 reads across all 50 samples were excluded from the analysis. All plotting figures show expression measured as reads per million reads (RPM) from normalised library sizes. Association of gene expression with age and sex was tested using generalised linear modelling (glm) implemented in *edgeR*, with age coded as a continuous variable and sex as a categorical variable. Preliminary analysis found

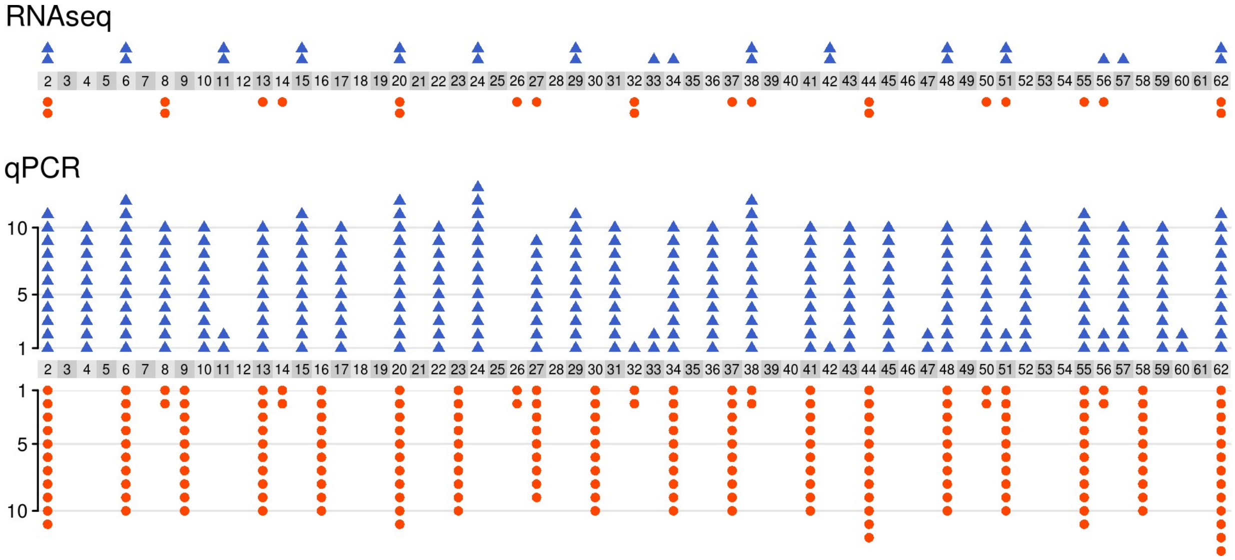


Fig 1. Number of samples used for RNAseq (top; total = 50) and qPCR (bottom; total = 498), split by age category (2–62 days old). Individual female and male flies shown as blue triangles and orange circles respectively.

https://doi.org/10.1371/journal.pntd.0009797.g001 little evidence of an important effect of the number of times a colony was exposed to the cold room on gene expression, but there was a significant effect of the number of days since flies had received a blood meal (S1 Text). We therefore controlled for days since receiving a blood meal by including it as a fixed continuous factor in the glm. False discovery rate control was set at 1% using the R package *fdrtool* [33].

Gene clustering analysis was performed with the *WGCNA* package in R [34], using the normalised read counts generated by e*dgeR* and keeping only the 5000 genes with the highest variance in expression. We used the hybrid module merging algorithm with a deep split value of 4, a minimum cluster size of 30 and a power parameter of 8, followed by module merging using the absolute value of the correlation coefficient between eigengenes as a distance matrix and a merging threshold of 0.2.

Prediction of age based on normalised read counts from the RNAseq data was performed using lasso regression implemented with the *glmnet* package in R [35]. As the aim was to find genes with consistently high predictive value for age, we explored a range of lasso parameters. This exploratory procedure is recorded in detail in the R script "02_lasso.r" provided on GitHub (https://github.com/EricRLucas/TsetseAgeMarkers).

2.4. Primer design and qPCR

Based on the results of the RNAseq analysis, 16 genes were short-listed to be tested as qPCR markers of age in *G*. *morsitans*, with two further genes being identified as suitable housekeeping genes for our purposes (i.e.: showed minimal variation in expression in the conditions included in our study and no evidence of association with age). Primers were designed for these genes based on the GmorY1.9 genome using NCBI Primer blast [36]. Where possible, amplicons were designed to span exon junctions. Based on testing amplification efficiency using 1:3 serial dilutions, the 10 best primer pairs for age-predictive genes, and the two primer pairs for housekeeping genes, were kept for use in the study and applied to 499 samples (298 females and 201 males), including 44 of the samples used for RNAseq (the remaining 6 samples had too little cDNA left to be included in the qPCR study). One of the samples failed to produce a Ct value for several genes and was therefore excluded from subsequent analysis, leaving 498 samples (Fig 1). All primers used in this study are listed in S2 Data.

qPCR was run on a AriaMX RealTime PCR instrument in a total volume of 20 μl, containing 10 μl of SYBR 2x MM, 1.2 μl of forward primer (5μM), 1.2 μl of reverse primer (5μM), 6.6 μl of nuclease-free water and 1 μl of genomic DNA. Reaction conditions: one cycle of 95˚C (3 minutes), 40 cycles of 95˚C (10 seconds) and 60˚C (10 seconds), one cycle of 95˚C (1 minute), 55˚C (30 seconds) and 95˚C (30 seconds, 5 seconds soak time).

Missing raw Ct values for age-predictive genes (where the signal never reached the threshold even after 40 cycles) were replaced with the maximum value of 40. ΔCt values were calculated using the mean Ct of the two housekeeping genes. Where Ct values were missing for either housekeeping gene, normalisation was impossible and the normalised aging gene value was recorded as missing (NA). All samples were run in two technical replicates and the final ΔCt was taken as the mean of the two replicates. Gene GMOY005321 consistently showed variable ΔCt values between technical replicates, possibly due to low expression of this gene, and these values were kept unchanged. For all other genes, any gene-sample combinations whose ΔCt differed by more than 1 between technical replicates were rerun for a third technical replicate, along with both housekeeping genes, providing a third ΔCt. In most cases, this third ΔCt was very close to one of the first two and very different from the other, indicating which of the first two technical replicates was wrong. The final ΔCt was thus taken as the mean of the third replicate and whichever of the first two replicates it was closest to.

2.5. Predicting tsetse age from qPCR data

Machine learning predictions of tsetse age from qPCR data were performed using the *caret* package in R (https://cran.r-project.org/package=caret). The ΔCt values for each of the 10 study genes were used as continuous predictor variables, and sex was included as a categorical predictor variable since some of the genes showed sex-dependent expression. Samples were randomly split into training set (75% of samples) and test set (25% of samples), stratified by sex and age to ensure equal representation of these two variables in the two sets. Due to rounding of sample numbers within each stratification layer, the final numbers in the train and test sets were 380 (76%) and 118 (24%) samples respectively. Model training was performed using three rounds of 10-fold cross-validation. For regression models, whose aim is to estimate age as a continuous variable, partial least squares regression (PLS), random forest and extreme gradient boosting (XGB) models were all trained on the data and their predictive accuracies compared. Categorical models were trained to categorise individuals into $\leq$15 and $>$15 days old. Simple decision tree, random forest and XGB models were compared for these categorical models.

The minimum number of expression markers (genes) required to obtain accurate predictions of age was determined by training the models with different numbers of loci. For each of the random forest and XGB models, the ten genes were ranked according to their variable importance in the full model training described above (sex was found to have a variable importance of 0 in both cases, and was therefore excluded from these models). The models were then trained with all ten genes, the top nine genes, the top eight genes, and so on. For each set of genes, 20 models were trained with a different random split of training and test sets, to account for stochastic variation in model accuracy.

All statistical analysis was conducted in R version 3.4.4 [37]. Analysis scripts are available on GitHub (https://github.com/EricRLucas/TsetseAgeMarkers).

3. Results

We collected 301 female and 204 male *G. morsitans* flies of known age from laboratory colonies, ranging in age from 2 to 62 days old. An initial RNAseq analysis of 28 female and 22 male samples showed that gene expression in these samples was primarily affected by age, rather than sex or days since last blood meal (Fig 2 and Fig A in S3 Text), although this was primarily due to the strong changes in gene expression found during the first 15 days of life, with older individuals clustering primarily by sex (Fig B in S3 text). Gene clustering analysis similarly showed that the largest cluster of correlated genes was one that changed strongly with age, particularly at young ages, with little effect of sex (S2 Text).

We identified a set of genes that was likely to provide strong age prediction by looking for genes that: 1. Were strongly correlated with age, or 2. consistently performed well in prediction of age using lasso regression and 3. where possible, belonged to different gene clusters as defined by weighted gene network clustering analysis. We particularly looked for genes showing strong expression changes in older individuals by identifying the genes most differentially expressed when considering only individuals older than 15 days, but even these showed relatively slight changes with age compared to some of the changes seen in the first 15 days of life (Fig 3 and Fig C in S3 Text). Using our criteria, and after testing qPCR primer efficiency, we manually picked 10 genes associated with age, and 2 genes with very little variation across samples to serve as housekeeping genes (Figs 3 and 4).

We obtained qPCR measurements of expression for these genes from 297 females and 201 males (Fig 1). As expected, expression of all 10 age-related genes was strongly correlated with age (Fig D in S3 Text) and with the RNAseq data (Fig E in S3 Text). Principal component

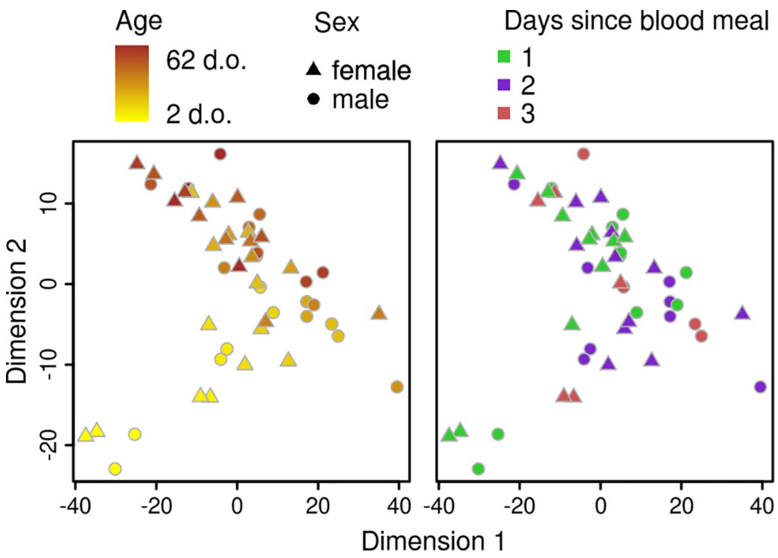


Fig 2. Gene expression clusters primarily by age. Principal component analysis of RNAseq data, coloured by age (left) or days since blood meal (right).

https://doi.org/10.1371/journal.pntd.0009797.g002

analysis of these age-related genes showed that age dominated the first principal component of the data. In particular, samples clustered strongly into those younger and older than 15 days (Fig F in S3 Text)

The qPCR expression data produced strong overall predictions of age, with predictions being much more accurate in young flies (15 days or younger) compared to older flies. For regression models, PLS provided the poorest predictions of age, while random forest and XGB models performed equally well (Fig 5 and Fig G in S3 Text). Taking the XGB model as an example, the overall root mean squared error (RMSE) for the final model was 6.74 days, but was 2.96 for individuals ≤15 days old. Variable importance for each gene in the random forest and XGB models are shown in Fig 4. Training the model separately for males and females did not improve prediction accuracy (Fig H in S3 Text).

Models also performed well at classifying samples into age categories of ≤15 and >15 days old (Fig I in S3 Text). The XGB model performed best in this task, accurately classifying 117 out of 118 samples in the test set.

For both the random forest and XGB regression models, prediction accuracy showed little decrease when the variables of least importance were dropped from the models (Fig 6). In both cases, accuracy remained comparable to that with all 10 genes when only 6 genes were included, with RMSE changing from 7.3 to 7.7 (random forest) or from 7.3 to 7.8 (XGB). In contrast, when moving to 5 genes instead of 6, RMSE changed from 7.7 to 8.4 (random forest) or from 7.8 to 9.3 (XGB). Interestingly, the same 6 genes proved to be sufficient for both model types (GMOY005321, GMOY002920, GMOY003090, GMOY003588, GMOY001603, GMOY003371). For the classification models, even fewer genes were needed (Fig 6), with just two genes being sufficient for XGB classification accuracy consistently better than 95% (GMOY002920, GMOY009908).

4. Discussion

We have identified a set of gene expression markers that can be used to predict the age of *G. morsitans* tsetse flies in the laboratory. Importantly, this method can be applied to both males and females, providing accurate estimates of age in male tsetse. This is particularly important

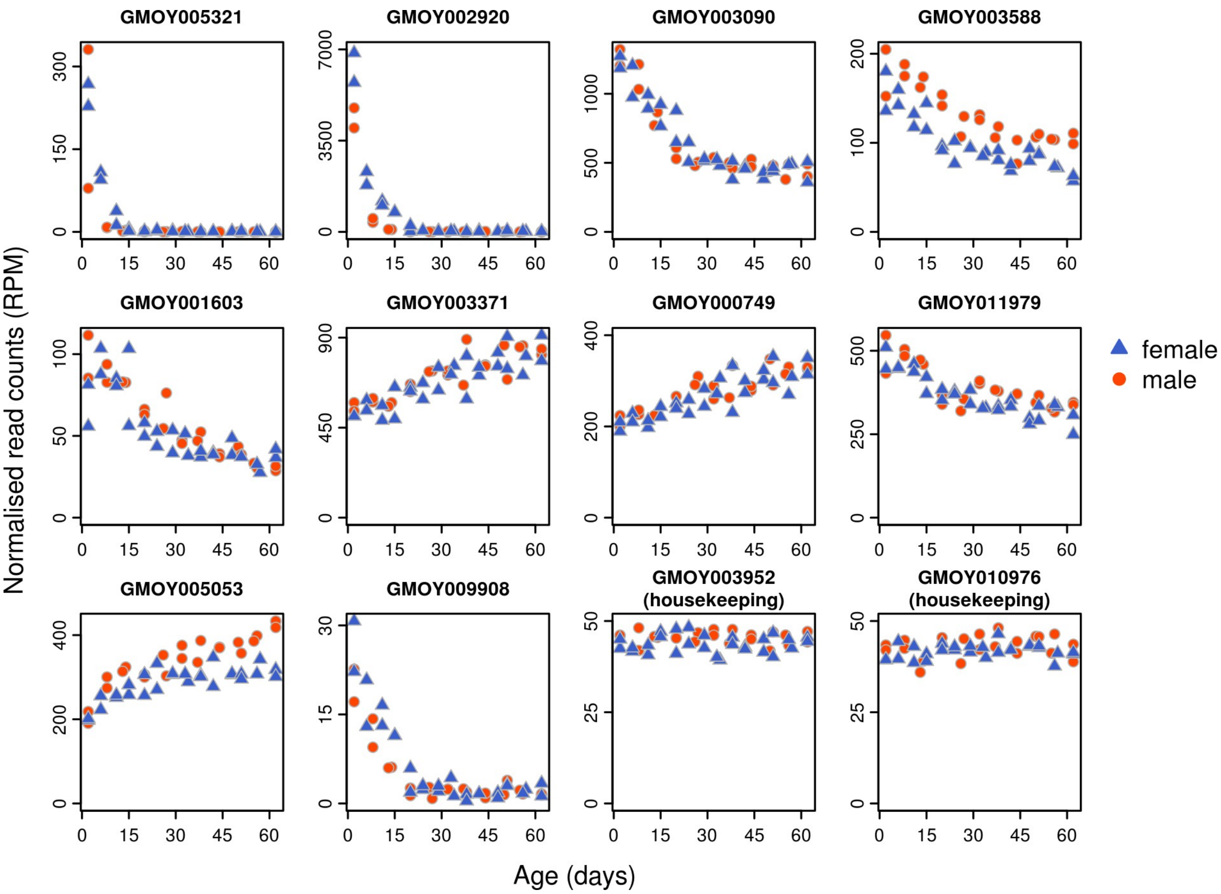
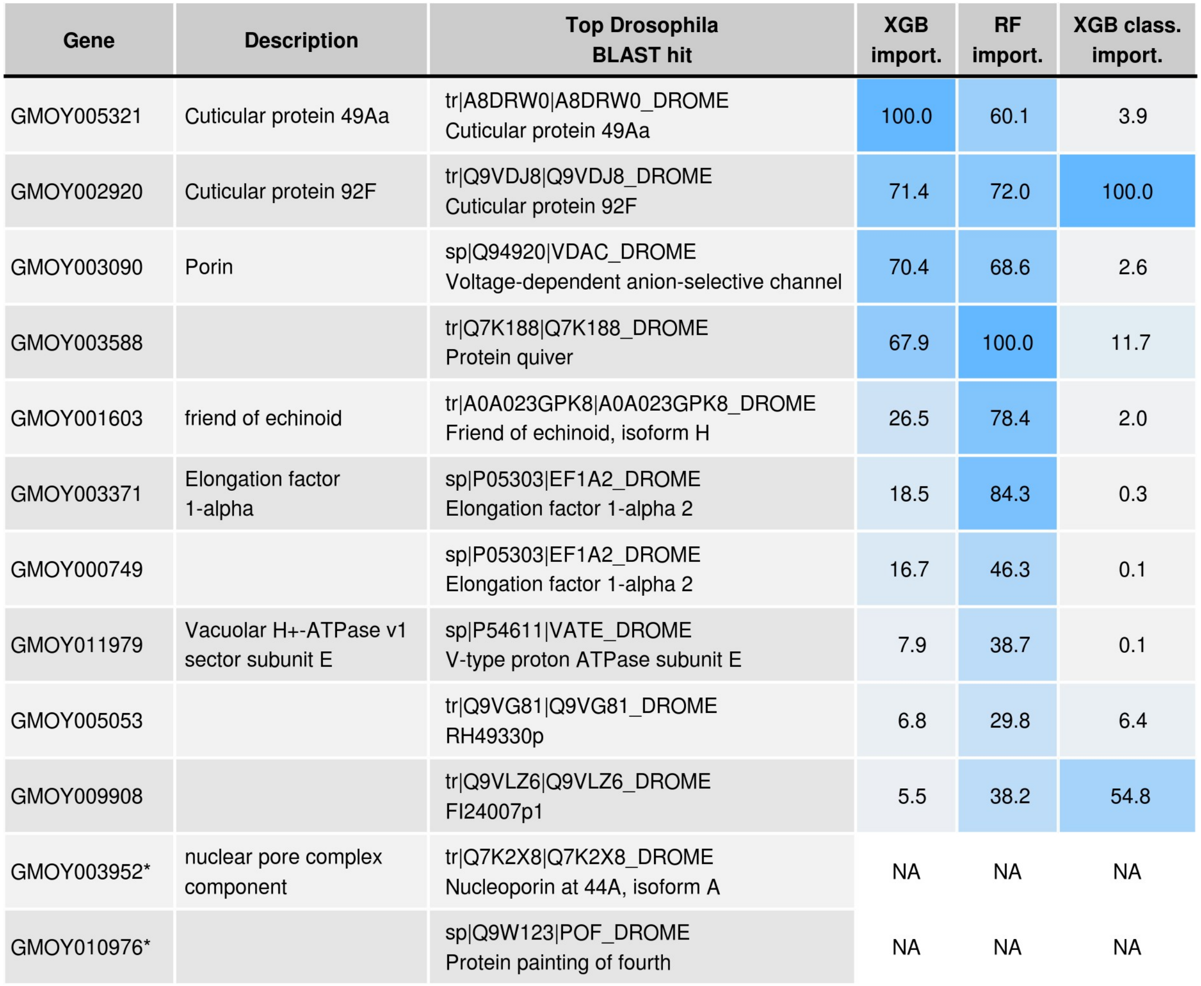


Fig 3. Expression of ten age-related genes and two housekeeping genes from RNAseq data, ordered according to the variable importance in the XGB model (Fig 4). Very strong early-age expression changes in some genes (eg: GMOY005321, GMOY002920) allow good discrimination among young individuals, but show little change in later life. Genes with continuous changes (eg: GMOY003371, GMOY000749) are more gradual and offer more consistent, but less powerful, discrimination at all ages.

https://doi.org/10.1371/journal.pntd.0009797.g003

since not only do both male and female tsetse flies transmit trypanosomes, but males appear to be more likely to develop transmissible infections [15,20]. Our genetic markers were also unaffected by time since an individual's last blood meal, making them more robust for use on wild-caught individuals, where such factors cannot be controlled.

Gene expression, like nearly all age-grading methods, is a measure of progression along some physiological trajectory. Thus, the predicted value is physiological age rather than chronological age, the former being dependent on the developmental rate of the individual while the latter is a strict measure of time elapsed since some defined notion of birth [38]. The correspondence between physiological and chronological age will thus depend on any factor that affects the developmental rate of the organism. In order to best identify markers that changed with age, we minimised the influence of such factors in our experiment by endeavouring to keep them fixed. For example, temperature and humidity were constant in our rearing conditions, and all samples were collected around the same time of day, leaving the possibility that these factors may yet influence the expression of our markers. Of course, in field conditions, these conditions will fluctuate, as will other potentially influential variables such as trypanosome infection, resource availability, circadian rhythms, stress and, in particular, seasonal and climatic fluctuations, which could affect expression directly or through their influence on developmental rate. Further work is required to test the applicability of the markers described



Gene	Description	Top Drosophila BLAST hit	XGB import.	RF import.	XGB class. import.
GMOY005321	Cuticular protein 49Aa	tr|A8DRW0|A8DRW0_DROME Cuticular protein 49Aa	100.0	60.1	3.9
GMOY002920	Cuticular protein 92F	tr|Q9VDJ8|Q9VDJ8_DROME Cuticular protein 92F	71.4	72.0	100.0
GMOY003090	Porin	sp|Q94920|VDAC_DROME Voltage-dependent anion-selective channel	70.4	68.6	2.6
GMOY003588		tr|Q7K188|Q7K188_DROME Protein quiver	67.9	100.0	11.7
GMOY001603	friend of echinoid	tr|A0A023GPK8|A0A023GPK8_DROME Friend of echinoid, isoform H	26.5	78.4	2.0
GMOY003371	Elongation factor 1-alpha	sp|P05303|EF1A2_DROME Elongation factor 1-alpha 2	18.5	84.3	0.3
GMOY000749		sp|P05303|EF1A2_DROME Elongation factor 1-alpha 2	16.7	46.3	0.1
GMOY011979	Vacuolar H+-ATPase v1 sector subunit E	sp|P54611|VATE_DROME V-type proton ATPase subunit E	7.9	38.7	0.1
GMOY005053		tr|Q9VG81|Q9VG81_DROME RH49330p	6.8	29.8	6.4
GMOY009908		tr|Q9VLZ6|Q9VLZ6_DROME FI24007p1	5.5	38.2	54.8
GMOY003952*	nuclear pore complex component	tr|Q7K2X8|Q7K2X8_DROME Nucleoporin at 44A, isoform A	NA	NA	NA
GMOY010976*		sp|Q9W123|POF_DROME Protein painting of fourth	NA	NA	NA

Fig 4. Ten age-related genes and two housekeeping genes (denoted with *) were used for qPCR analysis. Gene descriptions are taken from the Contig names in the GmorY1.9 proteome, downloaded from www.vectorbase.org/downloads on 2nd of March 2019. Top *Drosophila* BLAST hits obtained by blasting the GmorY1.9 proteome against the *D. melanogaster* swissprot proteome. Variable importance of each gene shown for XGB, random forest (RF) and XGB classifier models trained with all predictor variables.

https://doi.org/10.1371/journal.pntd.0009797.g004

here in field conditions. Before application as a monitoring tool, it will also be necessary to calibrate the model on a field population, and an important additional question to answer will be how valid that calibration is between field sites. Testing the accuracy of training the predictive model in one field site and applying it to other sites will indicate whether these markers can be used widely without re-calibration. If so, this will greatly improve the ease with which this tool can be applied.

The parameter perhaps most likely to affect the rate of development is environmental temperature. Mean daily temperatures fluctuate seasonally in the natural range of Glossina [21], and these fluctuations are known to affect developmental rate, for example decreasing the inter-larval period by around 0.35 days per degree [39]. If temperature or other environmental factors do affect the age-related change in expression, our markers will still be effective if the relationship between these factors and expression is determined, and the factors themselves are recorded over the study period. This would allow chronological age to be recovered from the expression-based estimates of physiological age. We also note that it is not yet clear whether chronological or physiological age is the most important parameter to estimate. For

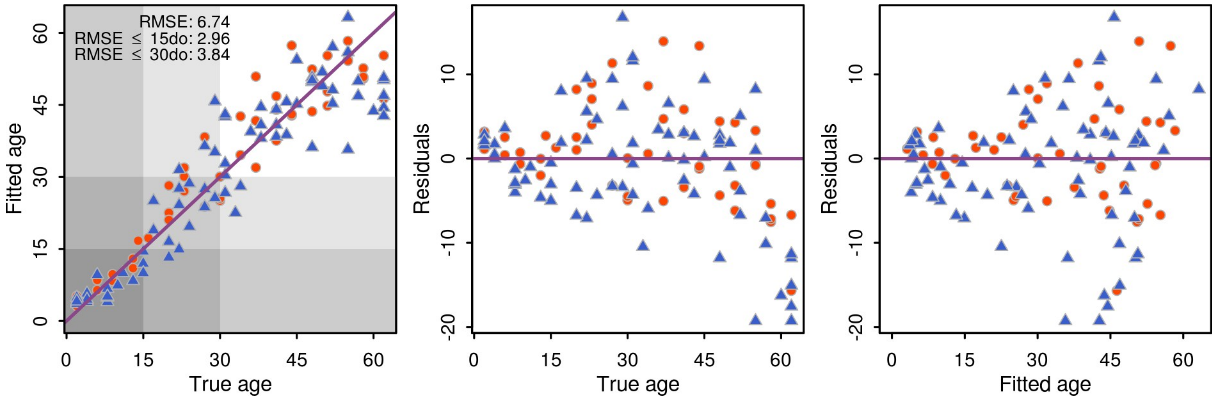


Fig 5. Prediction accuracy of the XGB model was highest (RMSE lowest) for individuals under 15 days old (2.96), and highest when all individuals were considered (6.74). Females are shown as blue triangles and males as orange circles. Purple line shows idealised perfect prediction.

https://doi.org/10.1371/journal.pntd.0009797.g005

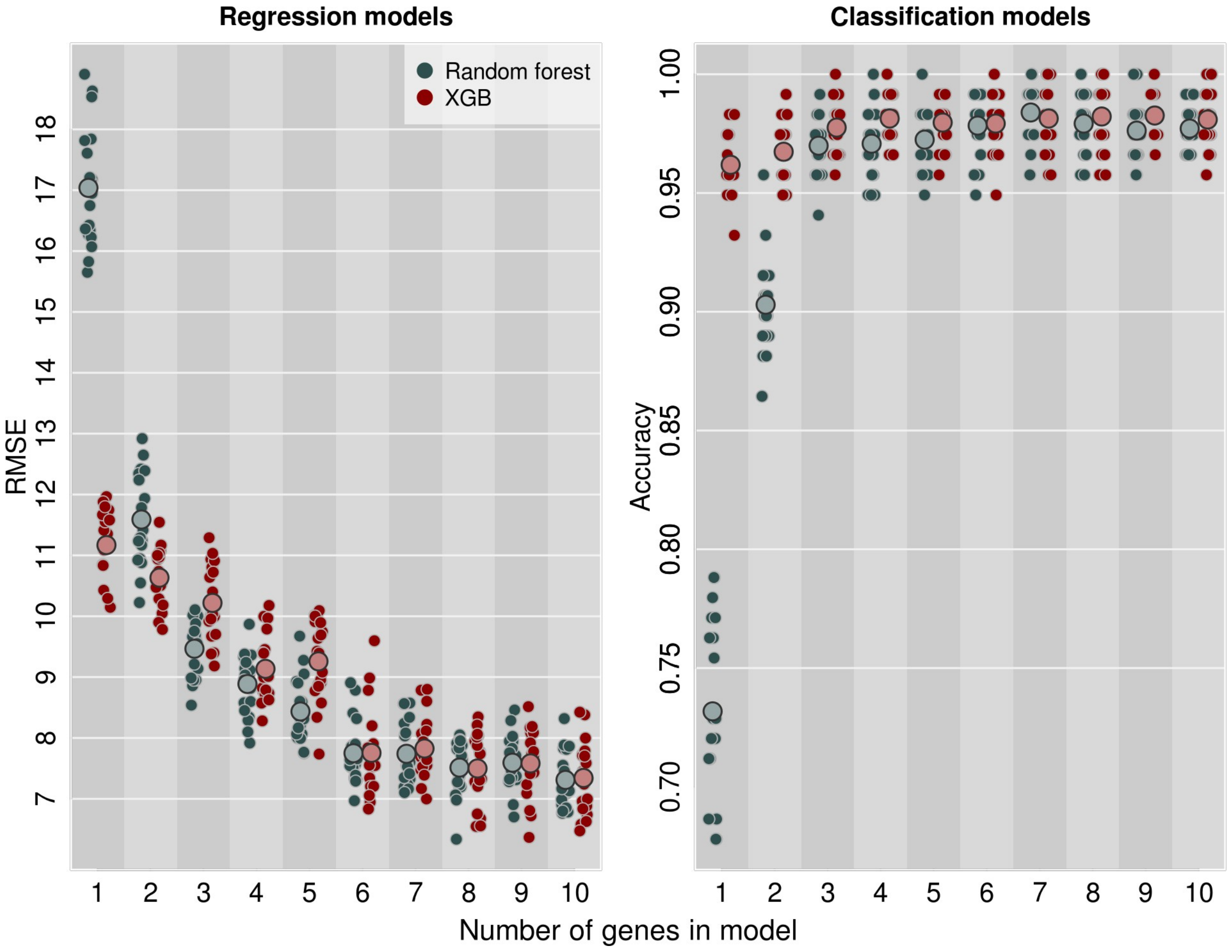


Fig 6. Predictive power of XGB and random forest models plateaus after the top 6 genes are included in the models (left). Accuracy of classification models plateaus after top 3 genes are included, with >95% accuracy achievable with only two genes (right). Small points show models run on independent test-train splits of the data (20 replicates per gene number); large points show the mean for each category. Points are jittered on the x axis to show overlapping data.

https://doi.org/10.1371/journal.pntd.0009797.g006 example, if increased developmental time leads to more frequent blood-feeding and / or shorter parasite maturation time, then flies will be infective at younger chronological age. Physiological age may then be the more important parameter.

Like other methods for estimating the age of vectors, prediction accuracy decreases at older ages [8,9,25,40–43]. In our data, this was because the change in expression with age was much greater in younger compared to older individuals, suggesting that the overall physiology of tsetse changes slowly after a certain life stage, and that there is thus little to detect that can be used for age grading. Although we found genes that continued to change in older ages, the rate of change relative to the variance within age groups was not sufficient to achieve the same prediction accuracies as found in younger individuals. While it is likely that more accurate old-age predictions would be achievable using whole-transcriptome methods such as RNAseq, this is too costly to be applied at the scales required for training predictive models. In mosquitoes, spectroscopy-based methods used to estimate age initially suffered from a similar loss of precision at older ages [9,43–45], but recent studies using machine learning prediction methods have improved prediction accuracies [46,47]. Whether similar performance can be achieved with tsetse should be explored.

The best estimates of maximum lifespan for *G. m. morsitans* in the field come from a mark-release-recapture study in which hundreds of newly-emerged tsetse were released and recapture attempts made over the course of six months [21,48]. Results indicated that around 30% of females survived for 60 days, and 10% survived for 110 days. In contrast, 10% of males survived to 30 days and 2% survived to 40 days. Thus, particularly for females, it would be advantageous to develop or improve age-grading tools that are accurate beyond the ages that have successfully been achieved so far.

While we used ten genes in our study, we found that using only the six genes most predictive of age still provided high prediction accuracy, and only two genes were needed for classifying individuals into age groups of $\leq$15 and $>$15 days old. By removing four genes from the analysis, qPCR time and costs can be reduced by 1/3 (eight qPCR reactions per sample instead of twelve), while removing eight genes will reduce costs by 2/3. We thus suggest that further studies testing the applicability of these markers in the field restrict themselves to either six or two genes, depending on how precisely age needs to be estimated. Such studies are needed to determine the applicability of these markers in the field, but it would also be interesting to measure the expression of these genes in age-controlled samples of other species of tsetse to determine whether these markers have widespread applicability. Once the field applicability of these markers is confirmed, the technique can be rolled out in the context of monitoring of tsetse control campaigns by comparing the age distribution before and after interventions to confirm that a resulting shift in the population age distribution is observed. In particular, in the wake of a 100% effective campaign, no flies older than the start of the campaign should be found. The resulting data on age structure both before and after control campaigns can then also be used to inform epidemiological models of trypanosomiasis transmission.

In conclusion, our study provides a new method for estimating the age of tsetse flies which does not require specialist dissection skills and can be applied to males. Testing of the applicability of these markers in the field in now required, and the problem remains of finding methods for more accurately estimating age in older individuals. This may involve identifying senescent changes whose rate is steady and consistent enough to be generalisable to any individual in the population.

Supporting information

S1 Data. Samples table.
(XLS)

S2 Data. Primer information table.
(XLS)

S1 Text. Preliminary RNAseq analysis details.
(PDF)

S2 Text. Gene clustering analysis results.
(PDF)

S3 Text. Supplementary figures. Fig A: Principle component analysis (PCA) of RNAseq data. Fig B: Hierarchical clustering of samples from RNAseq data reveal that young individuals ($<$ 15 days old) cluster together, with older individuals clustering by sex. Fig C: Expression changes with age for the 6 genes most strongly differentially expressed by age when only individuals older than 15 days were included in the model. Fig D: Correlation of qPCR measures of gene expression against age for the ten genes chosen as age markers. Fig E: Expression measured by qPCR and RNAseq were highly correlated in the samples in which both techniques were used. Fig F: PCA of samples based on qPCR measurements of expression of the 10 age-related genes. Fig G: Age prediction performance of PLS, random forest and XGB regression models. Fig H: Age prediction performance of random forest and XGB regression models trained separately on females and males. Fig I: Accuracy at classifying samples into age groups of $\leq$ 15 and $>$ 15 days old was 99% for the XGB classification model, 98% for the random forest and 97% for the decision tree.
(PDF)

Acknowledgments

We are grateful to Rob Leyland for assistance and tuition for the insectary work, and to Tom Churcher and Ben Lambert for valuable discussion on analytical approach.

Author Contributions

Conceptualization: Eric R. Lucas, Stephen J. Torr, Martin J. Donnelly.

Data curation: Eric R. Lucas, Alistair C. Darby.

Formal analysis: Eric R. Lucas.

Funding acquisition: Eric R. Lucas, Stephen J. Torr, Martin J. Donnelly.

Investigation: Eric R. Lucas.

Methodology: Eric R. Lucas.

Project administration: Eric R. Lucas.

Resources: Alistair C. Darby, Stephen J. Torr, Martin J. Donnelly.

Supervision: Stephen J. Torr, Martin J. Donnelly.

Visualization: Eric R. Lucas.

Writing – original draft: Eric R. Lucas.

Writing – review & editing: Alistair C. Darby, Stephen J. Torr, Martin J. Donnelly.

References

1. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018; 392: 1736–1788.

2. Eisler MC, Torr SJ, Coleman PG, Machila N, Morton JF. Integrated control of vector-borne diseases of livestock—pyrethroids: panacea or poison? Trends in Parasitology. 2003; 19: 341–345. https://doi.org/10.1016/s1471-4922(03)00164-8 PMID: 12901934

3. Shaw APM. Economics of African Trypanosomiasis. In: Maudlin I.; Holmes P. H. & Miles M. A. (editors). The Trypanosomiases. Wallingford: CABI Publishing; 2004. pp. 369–402.

4. Wilson AL, Courtenay O, Kelly-Hope LA, Scott TW, Takken W, Torr SJ, et al. The importance of vector control for the control and elimination of vector-borne diseases. PLoS Neglected Tropical Diseases. 2020; 14: e0007831. https://doi.org/10.1371/journal.pntd.0007831 PMID: 31945061

5. Geneva: World Health Organization. Monitoring and evaluation indicators for integrated vector management. 2012.

6. Wilson AL, Boelaert M, Kleinschmidt I, Pinder M, Scott TW, Tusting LS, et al. Evidence-based vector control? Improving the quality of vector control trials. Trends in parasitology. 2015; 31: 380–390. https://doi.org/10.1016/j.pt.2015.04.015 PMID: 25999026

7. Caragata EP, Poinsignon A, Moreira LA, Johnson PH, Leong YS, Ritchie SA, et al. Improved accuracy of the transcriptional profiling method of age grading in *Aedes aegypti* mosquitoes under laboratory and semi-field cage conditions and in the presence of *Wolbachia* infection. Insect molecular biology. 2011; 20: 215–224. https://doi.org/10.1111/j.1365-2583.2010.01059.x PMID: 21114562

8. Cook PE, Hugo LE, Iturbe-Ormaetxe I, Williams CR, Chenoweth SF, Ritchie SA, et al. The use of transcriptional profiles to predict adult mosquito age under field conditions. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103: 18060–18065. https://doi.org/10.1073/pnas.0604875103 PMID: 17110448

9. Sikulu M, Killeen GF, Hugo LE, Ryan PA, Dowell KM, Wirtz RA, et al. Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors. Parasites & vectors. 2010; 3: 49. https://doi.org/10.1186/1756-3305-3-49 PMID: 20525305

10. Dye C. The analysis of parasite transmission by bloodsucking insects. Annual review of entomology. 1992; 37: 1–19. https://doi.org/10.1146/annurev.en.37.010192.000245 PMID: 1539935

11. Woolhouse MEJ, Hargrove JW. On the interpretation of age-prevalence curves for trypanosome infections of tsetse flies. Parasitology. 1998; 116: 149–156. https://doi.org/10.1017/s0031182097002047 PMID: 9509024

12. Franco JR, Simarro PP, Diarra A, Ruiz-Postigo JA, Jannin JG. The journey towards elimination of gambiense human African trypanosomiasis: not far, nor easy. Parasitology. 2014; 141: 748–760. https://doi.org/10.1017/S0031182013002102 PMID: 24709291

13. Holmes P. Tsetse-transmitted trypanosomes-their biology, disease impact and control. Journal of invertebrate pathology. 2013; 112: S11–S14. https://doi.org/10.1016/j.jip.2012.07.014 PMID: 22841638

14. Rock KS, Torr SJ, Lumbala C, Keeling MJ. Quantitative evaluation of the strategy to eliminate human African trypanosomiasis in the Democratic Republic of Congo. Parasites & vectors. 2015; 8: 532. https://doi.org/10.1186/s13071-015-1131-8 PMID: 26490248

15. Dale C, Welburn SC, Maudlin I, Milligan PJM. The kinetics of maturation of trypanosome infections in tsetse. Parasitology. 1995; 111: 187–191. https://doi.org/10.1017/s0031182000064933 PMID: 7675533

16. Kgori P, Modo S, Torr S. The use of aerial spraying to eliminate tsetse from the Okavango Delta of Botswana. Acta Tropica. 2006; 99: 184–199. https://doi.org/10.1016/j.actatropica.2006.07.007 PMID: 16987491

17. Vale GA, Lovemore DF, Flint S, Cockbill GF. Odour-baited targets to control tsetse flies, Glossina spp. (Diptera: Glossinidae), in Zimbabwe. Bulletin of Entomological Research. 1988; 78: 31–49.

18. Van den Bossche P. The control of *Glossina morsitans morsitans* (Diptera: Glossinidae) in a settled area in Petauke District (Eastern Province, Zambia) using odour-baited targets. Onderstepoort Journal of Veterinary Research. 1997; 64: 251–257.

19. Hargrove JW. Age-specific changes in sperm levels among female tsetse (*Glossina* spp.) with a model for the time course of insemination. Physiological entomology. 2012; 37: 278–290.

20. Maudlin I, Welburn SC, Milligan P. Salivary gland infection: a sex-linked recessive character in tsetse? Acta Tropica. 1990; 48: 9–15. https://doi.org/10.1016/0001-706x(90)90060-d PMID: 1980807

21. Hargrove JW. Age-dependent changes in the probabilities of survival and capture of the tsetse, *Glossina morsitans morsitans* Westwood. International Journal of Tropical Insect Science. 1990; 11: 323–330.

22. Langley PA, Hall MJR, Felton T, Ceesay M. Determining the age of tsetse flies, *Glossina* spp.(Diptera: Glossinidae): an appraisal of the pteridine fluorescence technique. Bulletin of Entomological Research. 1988; 78: 387–395.

23. Lehane MJ, Hargrove J. Field experiments on a new method for determining age in tsetse flies (Diptera: Glossinidae). Ecological Entomology. 1988; 13: 319–322.

24. Lardeux F, UNG A, Chebret M. Spectrofluorometers are not adequate for aging *Aedes* and *Culex* (Diptera: Culicidae) using pteridine fluorescence. Journal of Medical Entomology. 2000; 37: 769–773. https://doi.org/10.1603/0022-2585-37.5.769 PMID: 11004793

25. Penilla RP, Rodríguez MH, López AD, Viader-Salvadó JM, Sánchez CN. Pteridine concentrations differ between insectary-reared and field-collected *Anopheles albimanus* mosquitoes of the same physiological age. Medical and veterinary entomology. 2002; 16: 225–234. https://doi.org/10.1046/j.1365-2915.2002.00364.x PMID: 12243223

26. Hargrove JW. A model for the relationship between wing fray and chronological and ovarian ages in tsetse (*Glossina* spp). Medical and Veterinary Entomology. 2020; 34: 251–263. https://doi.org/10.1111/mve.12439 PMID: 32222085

27. Rund SSC, Hou TY, Ward SM, Collins FH, Duffield GE. Genome-wide profiling of diel and circadian gene expression in the malaria vector *Anopheles gambiae*. Proceedings of the National Academy of Sciences. 2011; 108: E421–E430. https://doi.org/10.1073/pnas.1100584108 PMID: 21715657

28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal. 2011; 17: 10–12.

29. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010.

30. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

32. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010; 11: R25. https://doi.org/10.1186/gb-2010-11-3-r25 PMID: 20196867

33. Klaus B, Strimmer K. fdrtool: Estimation of (local) false discovery rates and higher Criticism. R package version 1.2.15. 2015.

34. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9: 559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008

35. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010; 33: 1. PMID: 20808728

36. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics. 2012; 13: 134. https://doi.org/10.1186/1471-2105-13-134 PMID: 22708584

37. R Core Team. R: A Language and Environment for Statistical Computing. 2015.

38. Hayes EJ, Wall R. Age-grading adult insects: a review of techniques. Physiological Entomology. 1999; 24: 1–10.

39. Hargrove J. Towards a general rule for estimating the stage of pregnancy in field-caught tsetse flies. Physiological Entomology. 1995; 20: 213–223.

40. Brei B, Edman JD, Gerade B, Clark JM. Relative abundance of two cuticular hydrocarbons indicates whether a mosquito is old enough to transmit malaria parasites. Journal of medical entomology. 2004; 41: 807–809. https://doi.org/10.1603/0022-2585-41.4.807 PMID: 15311480

41. Cook PE, Sinkins SP. Transcriptional profiling of *Anopheles gambiae* mosquitoes for adult age estimation. Insect molecular biology. 2010; 19: 745–751. https://doi.org/10.1111/j.1365-2583.2010.01034.x PMID: 20695922

42. Gerade BB, Lee SH, Scott TW, Edman JD, Harrington LC, Kitthawee S, et al. Field validation of *Aedes aegypti* (Diptera: Culicidae) age estimation by analysis of cuticular hydrocarbons. Journal of medical entomology. 2004; 41: 231–238. https://doi.org/10.1603/0022-2585-41.2.231 PMID: 15061283

43. Liebman K, Swamidoss I, Vizcaino L, Lenhart A, Dowell F, Wirtz R. The influence of diet on the use of near-infrared spectroscopy to determine the age of female *Aedes aegypti* mosquitoes. The American journal of tropical medicine and hygiene. 2015; 92: 1070–1075. https://doi.org/10.4269/ajtmh.14-0790 PMID: 25802436

44. Mayagaya VS, Michel K, Benedict MQ, Killeen GF, Wirtz RA, Ferguson HM, et al. Non-destructive determination of age and species of *Anopheles gambiae* sl using near-infrared spectroscopy. The American journal of tropical medicine and hygiene. 2009; 81: 622–630. https://doi.org/10.4269/ajtmh.2009.09-0192 PMID: 19815877

45. Sikulu-Lord MT, Milali MP, Henry M, Wirtz RA, Hugo LE, Dowell FE, et al. Near-Infrared Spectroscopy, a Rapid Method for Predicting the Age of Male and Female Wild-Type and *Wolbachia* Infected *Aedes aegypti*. PLoS Negl Trop Dis. 2016; 10: e0005040. https://doi.org/10.1371/journal.pntd.0005040 PMID: 27768689

46. Lambert B, Sikulu-Lord MT, Mayagaya VS, Devine G, Dowell F, Churcher TS. Monitoring the age of mosquito populations using near-infrared spectroscopy. Scientific reports. 2018; 8: 5274. https://doi.org/10.1038/s41598-018-22712-z PMID: 29588452

47. Milali MP, Sikulu-Lord MT, Kiware SS, Dowell FE, Corliss GF, Povinelli RJ. Age grading *An. gambiae* and *An. arabiensis* using near infrared spectra and artificial neural networks. PloS one. 2019; 14: e0209451. https://doi.org/10.1371/journal.pone.0209451 PMID: 31412028

48. Hargrove JW, Ouifki R, Ameh JE. A general model for mortality in adult tsetse (*Glossina* spp.). Medical and veterinary entomology. 2011; 25: 385–394. https://doi.org/10.1111/j.1365-2915.2011.00953.x PMID: 21414021