© 2023. Published by The Company of Biologists Ltd | Journal of Experimental Biology (2023) 226, jeb246548. doi:10.1242/jeb.246548

METHODS & TECHNIQUES

Scoring thermal limits in small insects using open-source, computer-assisted motion detection

Fernan R. Perez-Galvez[1,*], Sophia Zhou[1], Annabelle C. Wilson[1], Catherine L. Cornwell[2], David N. Awde[1,3] and Nicholas M. Teets[1]

ABSTRACT

Scoring thermal tolerance traits live or with recorded video can be time consuming and susceptible to observer bias, and as with many physiological measurements, there can be trade-offs between accuracy and throughput. Recent studies show that automated particle tracking is a viable alternative to manually scoring videos, although some of the software options are proprietary and costly. In this study, we present a novel strategy for automated scoring of thermal tolerance videos by inferring motor activity with motion detection using an open-source Python command line application called DIME (detector of insect motion endpoint). We apply our strategy to both dynamic and static thermal tolerance assays, and our results indicate that DIME can accurately measure thermal acclimation responses, generally agrees with visual estimates of thermal limits, and can significantly increase throughput over manual methods.

KEY WORDS: Automatic scoring, Thermal limits, Bioassay, Automated particle tracking, Motor performance

INTRODUCTION

Temperature influences nearly every aspect of an ectotherm's biology, which has fueled the measurement of thermal limits in a variety of organisms (Dallas and Rivers-Moore, 2012; Lutterschmidt and Hutchison, 1997b). Thermal limits are the minimum and maximum temperature at which a biological process can occur, and motor performance is perhaps the most used metric for assessing thermal limits. In insects, thermal tolerance provides physiological information directly related to fitness and is relevant for a number of research areas, ranging from basic ecophysiology (Addo-Bediako et al., 2000) to the impacts of climate change on insect diversity (Garcia-Robledo et al., 2016).

Thermal limits of motor performance can be scored using dynamic methods that involve increasing or decreasing the temperature until motor activity ceases, or, alternatively, insects can be exposed to a static, extreme thermal condition until motor failure occurs. Recent work indicates that dynamic and static thermal tolerance measures are mathematically, and perhaps physiologically, related (Jørgensen et al., 2019; Kingsolver and Umbanhowar, 2018; Rezende et al., 2014). Further, thermal tolerance can be assessed by measuring the resumption of activity after a period of paralysis, as is the case for the commonly used chill coma recovery time (Sinclair et al., 2015).

When an ectotherm approaches its thermal limits, it begins to be physiologically and behaviorally impaired. For critical thermal maxima (CT_{max}) or heat knockdown time (HKDT), the sequence of responses includes the loss of righting response, the sudden onset of muscular spasms, and finally the cessation of movement (Lutterschmidt and Hutchison, 1997b), which results in multiple possible interpretations of what constitutes the relevant physiological endpoint. While attempts have been made to standardize a measurement that would facilitate comparative analysis (Lutterschmidt and Hutchison, 1997a), the precise criteria can vary between phyla, making it difficult to compare results across studies (e.g. Sponsler and Appel, 1991; but see Sunday et al., 2011).

During chilling, a similar series of events occurs. When approaching the lower thermal limit, an ectotherm first slows or stops its normal activity, followed by a loss of coordination that impedes locomotion (i.e. the critical thermal minimum, CT_{min}), and finally, at lower temperatures, movement ceases altogether (i.e. chill coma onset) (Hazell and Bale, 2011). However, in practice, typically only the CT_{min} is reported and is often assessed by recording failure of a locomotor behavior (typically righting response, ability to cling to a surface, or a response to stimulus) (Sinclair et al., 2015). For chill coma recovery time (CCRT), recovery has been typically interpreted as the moment when the insect is 'able to stand on its legs' (David et al., 1998). Thus, there are many options available for assessing thermal tolerance, and clear, consistently applied endpoints are paramount for precision and repeatability.

In recent years, there has been an increase in large phenotypic screens to compare thermal tolerance across species (Kellermann et al., 2012; MacLean et al., 2019) or across genotypes of the same species (Gerken et al., 2015; Lecheta et al., 2020; Ørsted et al., 2018). Scoring thermal tolerance traits in real time or with recorded videos is time consuming, and there can be trade-offs between accuracy and throughput when analyzing large datasets. Thus, methods for automated scoring of thermal tolerance are necessary for improving repeatability and reducing strain on investigators.

Traditionally, thermal limits have been scored by monitoring individual insects in vials submerged in a water bath and observing insects in real time (Sinclair et al., 2015). Recently, Laursen et al. (2021) used particle tracking software (i.e. EthoVision XT) to score videos of thermal tolerance assays to increase automation and throughput. The results were qualitatively similar to manual estimates, although automated measurements were more variable as a result of visual artifacts and disturbances in the water bath, as well as the automated method's inability to detect subtle movements. Heating and cooling insects in air provides a means

[1]Department of Entomology, University of Kentucky, Lexington, KY 40508, USA. [2]Department of Mechanical Engineering, University of Kentucky, Lexington, KY 40508, USA. [3]Department of Ecology, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, 165 00 Praha, Czech Republic.

*Author for correspondence (fr_perezgalvez@outlook.com)

F.R.P.-G., 0000-0001-9620-9025

Received 17 August 2023; Accepted 11 October 2023

Journal of Experimental Biology (2023) 226, jeb246548. doi:10.1242/jeb.246548

to obtain higher quality video recordings, and MacLean et al. (2022) used a similar automated approach to score thermal tolerance videos of *Drosophila melanogaster* in acrylic arenas. Here, the automated method consistently recapitulated effect sizes in response to hardening treatments, and the differences in absolute values between the automated and traditional methods (i.e. in a water bath) are likely due to the differing thermal properties of the experimental apparatuses.

Thus, automated tracking methods appear to be a viable alternative to manually scoring videos, but continued efforts are needed to further benchmark these methods against classic human approaches and increase their flexibility for a variety of experiments that require assessment of motor activity. In particular, existing methods using particle tracking may not be suitable for traits that involve subtle movements such as rotation, limb movements and spasms, and some of the software options for particle tracking can be costly. Here, we present a novel strategy for automated scoring of thermal tolerance videos by inferring motor activity with motion detection. We tested three computational scoring methods and compared their results against visually obtained estimates to identify a computational interpretation of thermal limits that is most reliable. Our strategy is flexible, and we have applied it to both dynamic (CT_{max} and CT_{min}) and static (HKDT and CCRT) assays. Our method can accurately measure thermal acclimation responses, generally agrees with visual estimates of thermal limits, and can significantly increase the throughput over manual methods. We provide an open-source Python command line application we call DIME (detector of insect motion endpoint) that can be used to transform videos to motion data alongside R functions to estimate thermal limits using three distinct scoring methods.

MATERIALS AND METHODS

Thermal performance assays

Thermal performance was assessed in two dynamic assays (CT_{max}, CT_{min}) and two static assays (HKDT, CCRT) using the Oregon R strain of *Drosophila melanogaster* Meigen. Flies were reared at 25°C on a 12 h:12 h light:dark photoperiod on a standard cornmeal–yeast–molasses diet. To induce biological variation in thermal performance, adult flies were exposed to one of three acclimation treatments (18, 25 and 30°C) after adult emergence for a period of 5 days in programmable incubators (MIR-154, Panasonic Healthcare Co., Ltd). On the sixth day after emergence, flies were transferred to custom acrylic observation arenas using aspirators without anesthesia. Observation arenas hold up to 30 flies in individual wells (see README file in GitHub: https://github.com/fernan9/DIME) and were constructed from laser-cut acrylic layers. The wells in the arenas were sealed with a transparent acrylic lid on one side to facilitate recording, while the other side was sealed with nylon mesh to allow gas exchange. Flies that were crushed or mutilated during the loading process were removed from analyses.

Three replicates (blocks) of 30 individuals were performed per assay, giving a total sample size of 90 individuals. Each block included 5 males and 5 females from each of the three acclimation treatments, and the position of each sex-acclimation treatment combination was assigned randomly in the observation arena and was kept identical for every replicate. Observation arenas containing flies were placed in the center of a programable incubator (Panasonic Healthcare MIR-154) to perform thermal tolerance assays, which were recorded using a webcam (Logitech V-U0028). The setup was designed to keep the entire observation arena within the frame, avoiding light reflections of the light sources, and keeping a fixed lens focus to avoid spurious motion data due to variable depth of field. The distance between the objective and the observation arena varied between trials (12.5–19.0 cm), and recording distance did not affect the results.

In the case of the dynamic assays, CT_{min} and CT_{max}, a 0.25°C min^{-1} cooling or heating ramp starting at 25°C was programmed in the incubator with a function that changes temperature by 2.5°C every 10 min interval. Even though the program includes discrete temperature steps, the heating and cooling capacity of the incubator, coupled with thermal buffering by the arena, led to an approximately linear thermal ramp (coefficient of determination R^2 CT_{min}: 0.999, 0.998, 0.999; R^2 CT_{max}: 0.999, 0.999, 0.986), as measured in the well microenvironment with a DHT22 temperature sensor (Aosong Electronics Co., Ltd) and an Arduino Nano microcontroller platform (Arduino SRL).

After the experiment, a linear model was fitted to the cooling or heating section of the ramp, and the linear coefficients were used to translate the time of knockdown into a CT_{min} or CT_{max}. The slopes measured in both cases were constant throughout the duration of the assay but slightly less steep than programmed (CT_{min}: −0.227, −0.223, −0.230°C min^{-1}; CT_{max}: 0.236, 0.238, 0.230°C min^{-1}). In static assays, a constant temperature of 36.5°C was used for HKDT. Two incubators were used to perform CCRT bioassays; the first was used to induce chill coma for 2 h at 0°C followed by immediate transfer to the second incubator for recovery at 25°C. Examples of the thermal profile for each assay are given in Fig. S1, with slopes indicated in the case of dynamic assays and an approximate time to reach 90% of the target temperature in the case of static assays.

Transformation of insect motor activity

Our command line application DIME transformed biological activity in thermal performance videos to a numerical variable of motion detection. DIME was developed in Python v3.8.8 and makes use of the computer vision library OpenCV v4.5.3 (Bradski, 2000). The program transforms motion to a numerical variable which is a measurement of relative pixel intensity change (rPIC) within a region of interest (ROI). Each ROI contains a single individual insect on a constant background and must be drawn by hand at the beginning of the computer analysis (see README file in GitHub: https://github.com/fernan9/DIME).

A video is analyzed as a series of images representing single frames of video data (video frame), and the number of images extracted per second depends on the frame rate of the video, which is extracted automatically from the metadata of the file at the beginning of the analysis. Video frames are extracted as pixel matrices in the blue–green–red (BGR) color space and transformed to a single grayscale value using the standard-definition luminance formula $Y'=0.299R+0.587G+0.114B$ as implemented in the OpenCV command COLOR_BGR2GRAY. The procedure continues by computing the difference between pairs of consecutive grayscale pixel matrices to generate a series of difference matrices. Motion between video frames is encoded on each pixel of the difference matrix as a deviation from 0. Three filters are applied to each difference matrix, a Gaussian blur to remove flickering particles in the background using the OpenCV command GaussianBlur (vertical and horizonal kernel size [k]=3), a dilation filter to maximize the difference between areas of change with the command dilate (convolution iterations [i]=2), and a binary threshold to set pixel values to either complete white (0) or black (255) using a threshold of 20. Finally, motion detection per individual is achieved by calculating the average absolute pixel intensity change per ROI in the filtered difference matrix. As the

Journal of Experimental Biology (2023) 226, jeb246548. doi:10.1242/jeb.246548

maximum value of intensity in pixel change is 255 (complete black), the values of rPIC will vary from 0 when no motion is detected in the ROI to 255 when every pixel changes.

An additional subtraction filter was applied to rPIC data to remove spurious movement originating from strong vibration in the incubator, automatic refocusing of the lens or background movement. The filter changes the rPIC values for an entire frame difference that is higher than a user-selected threshold to zero. This threshold can be selected from a histogram that is plotted after each run and must be applied independently per video file. Extreme peaks must be identified, and multiple iterations may be necessary to achieve the desired result. Application of the filter in this study and a short description of the procedure of filtering are available from GitHub (https://github.com/fernan9/DIME).

Inference of thermal tolerance endpoints

Bioassay scoring has been traditionally performed by human observers who are trained to identify and record behavioral changes which are later analyzed and interpreted. In this study, we prompted experimenters to score the endpoint for CT_{min}, CT_{max} and HKDT as the time when the individual moved to the last position in the observation well, while for CCRT, the endpoint was scored as the moment when the fly recovered an upright position. With experience, the experimenter's records become more accurate as they learnt to discard behaviors that are misleading and optimize the recording process using their own methodologies. To test for interobserver variation, we assigned three experimenters to score a subset of the dataset and compared the precision of estimates between observers and against the computational methods (see below).

For computational assessment of thermal tolerance, we compared three computational methods to extract endpoints from the motion data: change point, individual median and optimal threshold. The change point method takes advantage of the large amount of motion data collected during the assay and applies a statistical model to identify the time point where activity starts or ends. The individual median and optimal threshold methods are computational approaches mimicking the heuristic applied by experimenters to score videos: panning through the video from inactivity to activity and determining the first movement observed either as the activity onset or the endpoint. Comparing these approaches would provide information on the bias introduced by independent scoring methodologies.

Specifically, the change point method uses the entire activity data to statistically determine the time point where activity changes between active and inactive states. In this interpretation, the change point method considers motor activity in rPIC as a sequence of observations with an underlying pattern where the initial and final means are different, and the change point between them is unknown (Hinkley, 1970). Here, we apply a maximum likelihood estimator to identify a single change point of motion along the thermal performance assay of each experimental subject using the 'at most one change' command as implemented in the R library changepoint v2.2.3 (Killick and Eckley, 2014).

Both the individual median and optimal threshold methods use a threshold to determine which rPIC values reflect motor activity, but the methods differ in the way the threshold is determined. The threshold for individual median is determined using the median of the non-zero rPIC values observed individually per well. The optimal threshold method is a modified version of the scoring methodology described in MacLean et al. (2022) where the threshold is determined per video and is optimized using the data of all individuals. In their definition, the algorithm first identifies the maximum activity level of uninformative data (noise) and then scores the last motion event above this threshold (MacLean et al., 2022).

Based on code provided in MacLean et al. (2022), we implemented their algorithm as the optimal threshold method by using a series of n thresholds (default n=10). As most of the variation is usually present in the lower distribution of the data, the threshold values are evenly distributed between 0 and 70% of the maximum activity recorded in the first individual. Estimates for every individual are computed for each threshold, providing n datasets. Consecutive pairs of datasets with increasing threshold values are then fitted to a linear regression, providing $n-1$ regression lines (Fig. S1). The first regression with the highest R^2 is said to be where the scoring becomes stable (i.e. when all individuals are scored above the noise level). The optimal threshold is the one with the minimum value between the pair used for the selected regression. For CT_{max}, CT_{min} and HKDT, the last value above this threshold is considered the endpoint of activity and is scored as the thermal limit. In the case of CCRT, the algorithm is applied to the reversed data to capture the first event. One modification was made to the MacLean et al. (2022) method: originally the last event was scored only on decreasing values as they approach the noise threshold; we removed this condition and allowed any value to be scored, as ending motions such as a spasm or a last jump may be larger in magnitude than immediately previous events. Despite this difference, we expect the original and our modified version of optimal threshold to have similar performance.

Analysis of computational scoring reliability

Each experimental block was scored visually by one of the authors (block 1: F.R.P.-G., block 2: A.C.W., block 3: S.Z.), who each had different experience scoring thermal limits to simulate a realistic large-screening experimental setup. All statistical analyses were conducted using R v4.1.0 (http://www.R-project.org/). First, an exploratory analysis of variance (ANOVA) was conducted in joint datasets containing the visual and computational estimates to identify variance associated with the methodology, using the model:

$$\text{Thermal limit estimate} \sim \text{Temperature}_{\text{acclimation}} + \text{Method} + \text{Sex} + \text{Block}. \quad (1)$$

A *post hoc* Tukey test for honestly significant differences (HSD) was applied to identify significant average differences between computational and visual methodologies. No significant variance associated with the variable Sex was detected in lower thermal tolerance assays (CT_{min} and CCRT), so we decided to exclude this variable from the rest of the analyses. Also, significant variance associated with the variable Block was observed (CT_{max}: $F_{2,348}$=48.33, P≤0.001; CT_{min}: $F_{2,226}$=13.99, P≤0.001; HKDT: $F_{2,339}$=76.98, P≤0.001; CCRT: $F_{2,256}$=7.8, P≤0.001). With this information, we fitted datasets from independent scoring methods to a mixed effects model using acclimation temperature ($\text{Temperature}_{\text{acclimation}}$) as a fixed effect and Block as random effect with the R library lmer from the lme4 v1.1-29 package (Bates et al., 2015), using the equation:

$$\text{Thermal limit} \sim \text{Temperature}_{\text{acclimation}} + (1|\text{Block}). \quad (2)$$

Population marginal means, their associated standard errors, and the *post hoc* Tukey HSD test applied to pairwise differences between treatment levels were calculated using the package emmeans v1.8.0

(https://CRAN.R-project.org/package=emmeans). The Kenward–Roger approximation of degrees of freedom was applied when testing independent scoring methods to account for small and unbalanced datasets (e.g. when a methodology was not able to provide an estimate), the confidence intervals were adjusted using the Šidák method, and the *P*-values were adjusted using the Tukey method for comparing a family of three estimates.

Inter-method agreement between computational and visual estimates was measured with the concordance correlation coefficient (CCC), which evaluates the degree to which pairs of measurements fall in the 45 deg line of perfect correlation (Lawrence and Lin, 1989). In addition, CCC can provide information on the source of disagreement when decomposed into the bias corrector factor C_b, a measure of accuracy, and the Pearson correlation coefficient ρ, a measure of precision. Accuracy in C_b measures how far the best-fit line deviates from the 45 deg line; ρ measures how far each observation deviates from the best-fit line. In our case, CCC and their components were computed as implemented in the R package DescTools v0.99.45 (https://CRAN.R-project.org/package=DescTools) in paired datasets containing one computational method (change point, individual median, optimal threshold) and the visual dataset per thermal tolerance assay. Finally, individuals presenting outlying scoring differences were identified with an agreement test (Martin Bland and Altman, 1986) to investigate the cause of the disagreement. For each thermal tolerance assay, we calculated the mean difference ($\bar{d}$) and the standard deviation of the difference in the comparisons visual–individual median and visual–optimal threshold datasets to estimate the 'limits of agreement' at $\bar{d}$±2 s.d.

Application to other insect species

The CT_{max} of six additional insect species was evaluated using our methodology (individual median) and visual estimations. The additional species tested were the Asian tiger mosquito (*Aedes albopictus*), the southern house mosquito (*Culex quinquefasciatus*), the common bed bug (*Cimex lectularius*), the subterranean termite (*Reticulitermes flavipes*), the fall army worm (*Spodoptera frugiperda*) and the red flour beetle (*Tribolium castaneum*). Insects were reared under standard conditions at multiple insectariums in the University of Kentucky Department of

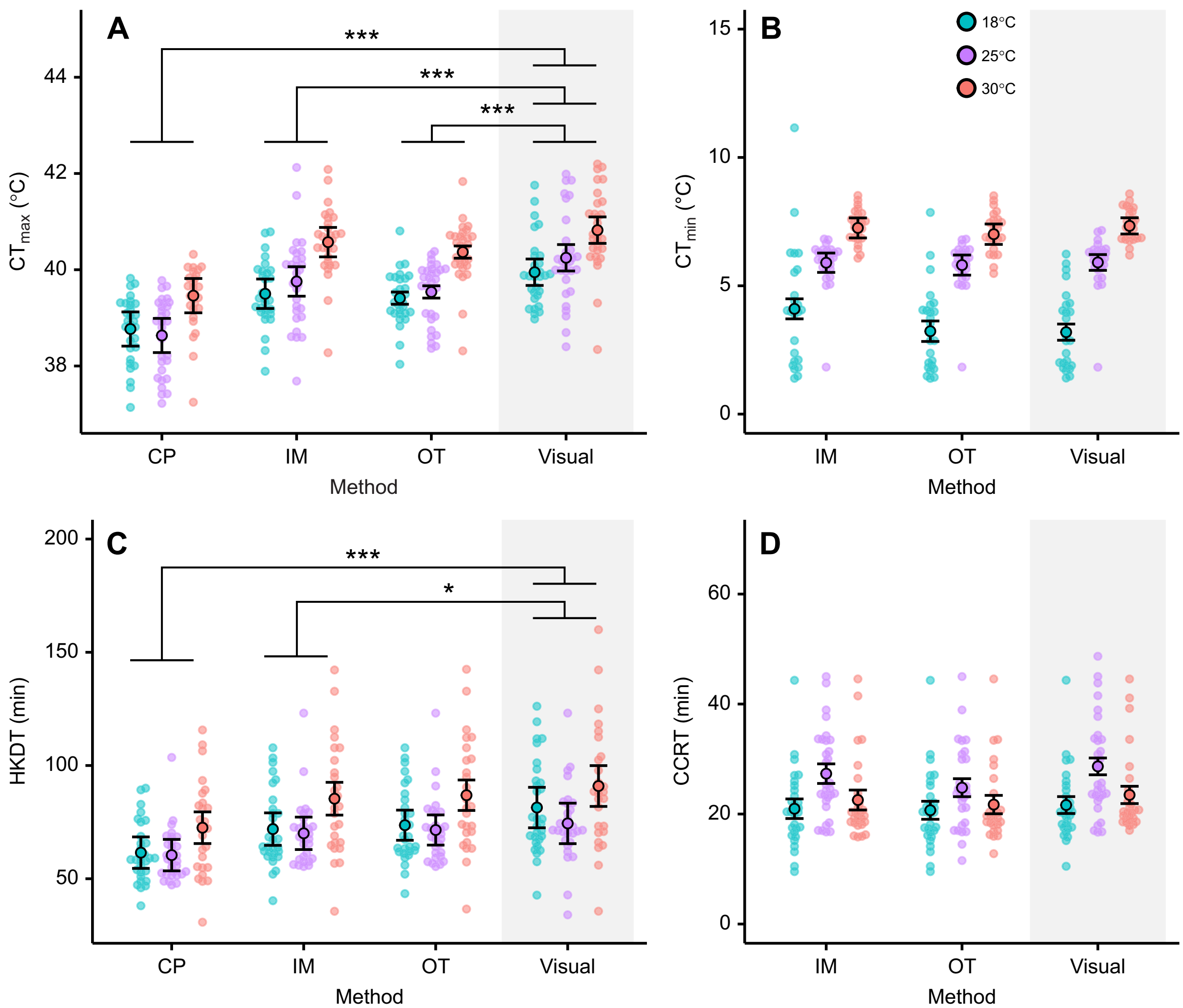

Fig. 1. Comparison between thermal tolerance estimates scored with different methodologies. The methodologies were change point (CP), individual median (IM) and optimized threshold (OT) versus visual estimates. (A) Critical thermal maxima (CT_{max}; *n*=90 for IM, OT and visual; *n*=87 for CP). (B) Critical thermal minima (CT_{min}; *n*=78 for IM, OT and visual). (C) Heat knockdown time (HKDT; *n*=87 for all). (D) Chill coma recovery time (CCRT; *n*=88 for IM, OT and visual). Colored circles are raw data, outlined circles represent mean treatment values and error bars are s.e.m. for the three different acclimation temperatures. Asterisks indicate statistical significance with Tukey's HSD test (**P*<0.05, ****P*<0.001).

Journal of Experimental Biology (2023) 226, jeb246548. doi:10.1242/jeb.246548

Entomology and loaded onto acrylic plates for experiments immediately after receipt, except for fall army worm larvae, which were held in our laboratory until the 3rd instar was reached, and common bedbugs, which were held for 7 days at room temperature after a blood meal. When available, a previously determined estimate was obtained from the literature. The incubator was programmed for a 20 min holding time at 25°C followed by the 0.25°C min^{-1} heating ramp, except for fall army worm larvae, for which it was held at 30°C before starting a ramp with the same rate.

RESULTS AND DISCUSSION

Scoring efficacy by computational methods

Motor activity was transformed into a sequence of motion events with our computational tool DIME. Average processing time including transformation and scoring was 18 min (1.80 GHz processor) per 2.5 h long video (640×360 pixels) in contrast to 1 h for visual inspection of the same 30 subjects. As processing time increases with increasing pixel resolution, video frame manipulation takes up most of the computing resources. Greater throughput can be achieved by increasing the number of subjects per video and increased processing power. The computational methods individual median and optimal threshold provided a score for every individual; however, for lower thermal limits, change point scores were removed from the downstream analysis as most were not meaningful.

Variance introduced by methodology

Variation due to methodology was only observed in upper thermal limits. CT_{max} presented the greatest differences between automated and visual scoring methods (Fig. 1A). Both the individual median and optimal threshold methods resulted in statistically significant deviations from visual estimations (−0.39 and −0.56°C, respectively, both P<0.001), whilst the magnitude of the difference with the change point method was more pronounced (−1.39°C, P<0.001). Methodological differences were also present in HKDT assays (Fig. 1C). Estimates obtained using the change point and individual median methods differed from visual observations on average by −17.3 min (P<0.001) and −6.48 min (P=0.03), respectively, while optimal threshold estimates were not significantly different (−4.88 min, P=0.14). In general, estimates from the change point method were smaller than those using the individual median and optimal threshold methods, suggesting that this approach may be scoring a different component of the biological response to thermal stress.

In contrast, no statistical variance was introduced by methodology in lower thermal limits. In the case of CT_{min}, the average differences of individual median and optimal threshold against visual estimates were 0.26°C (P=0.34) and −0.11°C (P=0.83). In a similar way, CCRT estimates between computational and visual estimates were −0.97 min (P=0.64) and −2.2 min (P=0.1), respectively. The similarity in variances between methods could be a result of the reduced amount of motion events, but the relative lack of movement may also introduce variation when automatically scoring some individuals (see sources of disagreement below).

Recapitulation of thermal acclimation effects

In dynamic assays, computational scoring methodologies consistently recapitulated the differences in thermal tolerance due to acclimation (Table 1). In contrast, in static assays, there was some variation in statistical groupings between scoring methodologies. In visual estimates of HKDT, flies acclimated at 25°C were statistically different from those acclimated at 30°C, while flies acclimated at 18°C were not different from either group (ANOVA, Tukey, α=0.05). For automated scoring, statistical grouping was slightly different, as the 18 and 25°C acclimation groups were indistinguishable and both were different from the 30°C acclimation group. The reduced s.e.m. in automatic scores suggests that these methodologies had increased power to identify treatment effects by providing reduced measurement error, supporting the notion that interobserver bias could be a source of reduced statistical power in HKDT (Castaneda et al., 2012). In the case of CCRT, treatment grouping from the visual estimates was recapitulated by individual median scoring, but not by optimal threshold scoring. In this case, the optimal threshold method was not able to separate treatments from each other, despite having lower s.e.m. than individual median estimates. However, on the whole, our observations were consistent with previous observations that automatic scores recapitulate treatment effects (Laursen et al., 2021; MacLean et al., 2022), confirming that automated scores can provide meaningful thermal limit estimates.

Concordance between automated and visual estimates

To evaluate the reliability of automated scoring, we used the CCC, a measure of agreement. The decomposition of CCC into C_b and ρ provides specific information on accuracy with the bias correction factor (C_b) and precision with the product-moment correlation (ρ) (Table S1). The agreement of automatic scoring methodologies was higher in static than in dynamic methods, with individual median providing an overall accuracy above 91% but variable precision (range 66–91%), and optimal threshold presenting a similar pattern (>79% C_b, 50–92% ρ). Scores from the change point method presented the lowest CCC values, confirming our suspicion that this scoring method could be measuring a different component of thermal performance, albeit with high correlation (ρ) for HKDT. Our CCC values for automated versus visual estimates were comparable to or even higher than previously reported for HKDT (Castaneda et al., 2012) but slightly lower than those obtained by our interobserver CCC estimates (Table S1). In general, the capacity of individual

Table 1. Effect of temperature on thermal tolerance estimates of CT_{max}, CT_{min}, HKDT and CCRT calculated with change point, individual median, optimal threshold and visual methods

	Acclimation temperature		
Method	18°C	25°C	30°C
CT_{max} (°C)			
Change point	38.6±0.37^a	38.7±0.37^a	39.4±0.371^b
Individual median	39.4±0.27^a	39.6±0.27^a	40.4±0.27^b
Optimal threshold	39.4±0.15^a	39.5±0.15^a	40.4±0.15^b
Visual	39.9±0.274^a	40.2±0.274^a	40.8±0.274^b
CT_{min} (°C)			
Individual median	4.1±0.39^a	5.9±0.377^b	7.25±0.39^c
Optimal threshold	3.23±0.398^a	5.81±0.391^b	7.01±0.398^c
Visual	3.19±0.316^a	5.91±0.307^b	7.33±0.316^c
HKDT (min)			
Change point	61.5±6.97^a	60.4±6.97^a	71.9±6.99^b
Individual median	71.9±7.2^a	70.1±7.2^a	84.4±7.24^b
Optimal threshold	73.6±6.72^a	71.5±6.72^a	85.7±6.77^b
Visual	81.4±9.04^{a,b}	74.4±9.04^a	88.2±9.09^b
CCRT (min)			
Individual median	21±1.78^a	22.5±1.82^a	27.3±1.78^b
Optimal threshold	20.7±1.64^a	21.7±1.67^a	24.8±1.64^a
Visual	21.6±1.53^a	23.5±1.58^a	28.7±1.53^b

CT_{max}, critical thermal maximum; CT_{min}, critical thermal minimum; HKDT, heat knockdown time; CCRT, chill coma recovery time. Data are means±s.e.m. Means not sharing any letter are significantly different (Tukey's HSD test at the 5% level of significance).

Journal of Experimental Biology (2023) 226, jeb246548. doi:10.1242/jeb.246548

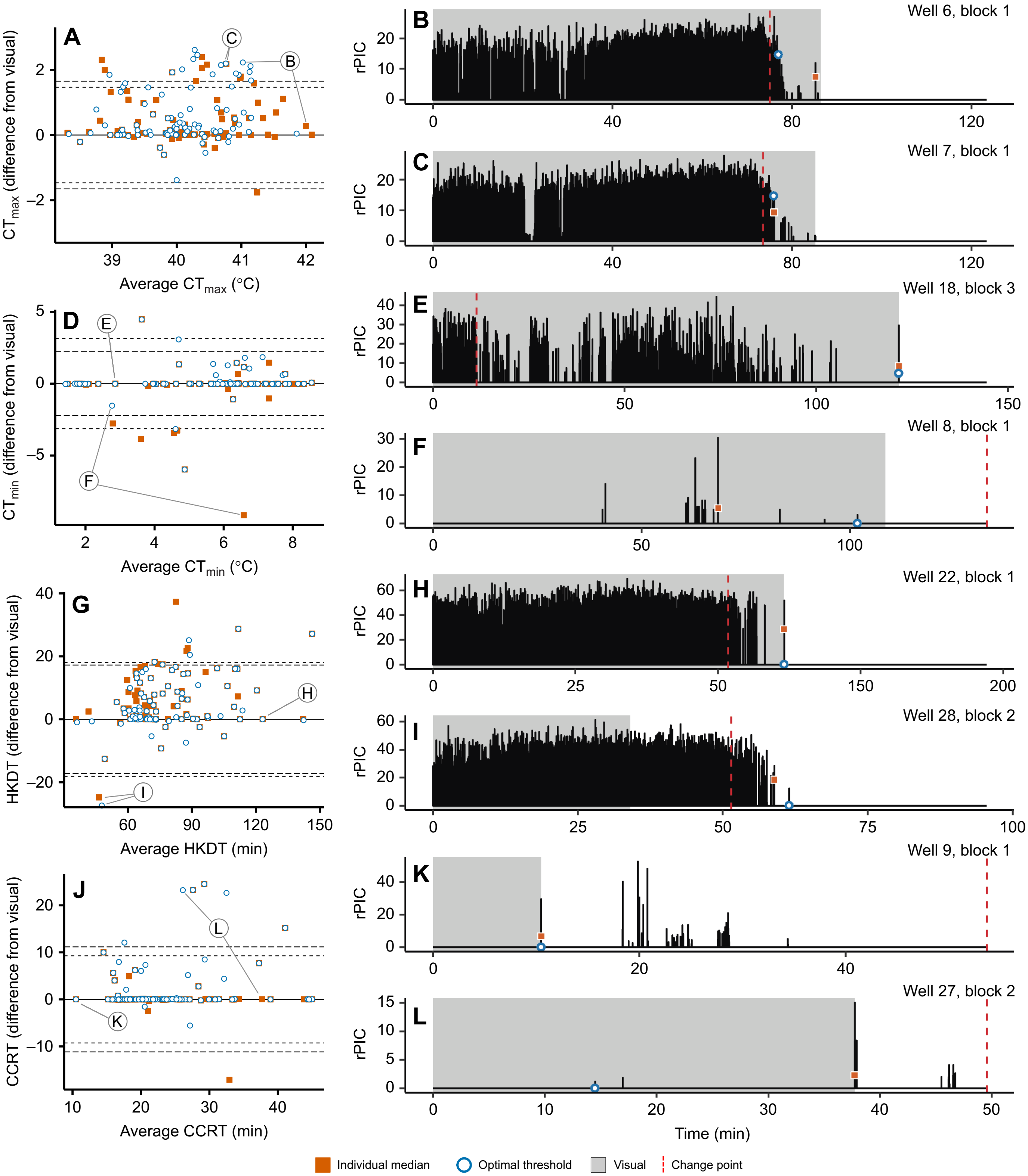

Fig. 2. Source of discrepancies between pairs of measurements. Agreement tests (left) for individual median and threshold optimization against visual scores, and two individual examples of each assay (right) for (A–C) CT_{max} (n=90), (D–F) CT_{min} (n=78), (G–I) HKDT (n=87) and (J–L) CCRT (n=88). In the agreement tests, short-dashed lines indicate the limits of agreement based on individual median–visual differences, and long-dashed lines are for optimal threshold–visual differences. Orange squares indicate differences for scores estimated with the individual median method; blue circles indicate differences for those estimated with the optimal threshold method. In the activity plots on the right (showing relative pixel intensity change, rPIC), the time where the gray shading ends indicates the visual estimate, and the red vertical dashed line indicates the estimate obtained with the change point method.

median and optimal threshold methods to accurately capture the thermal limits is supported by the reduced bias (high C_b), but the variable ρ suggests that discrepancies exist between individual pairs of measurements.

Source of disagreement between pairs of measurements

To identify the source of the disagreement between individual scorings, we evaluated individual median and optimal threshold datasets against visual estimates using a graphical technique (Fig. 2). In many cases, the scores from visual, individual median and optimal threshold methods overlapped (e.g. Fig. 2E,H,K), particularly in CT_{min} and CCRT datasets. As expected from the CCC, greater individual variation between visual and automated individual scoring was present in CT_{max} and HKDT. The majority of the discrepancies were underestimations of the endpoint time by the automated method, originating from an interaction between the decreasing activity levels in the last movements of the individual and the threshold differences between scoring methods (e.g. Fig. 2B,C).

Biological variation in the motor response also introduced technical error. For example, individuals with slow movements were not detected accurately by the motion detection software, as in the case of well 8 in block 1 of CT_{min} (Fig. 2F). In other cases when the scoring threshold was low, automated CCRT estimates were scored on small appendage motion events instead of righting position because of its high sensitivity (e.g. optimal threshold estimate in Fig. 2L). These discrepancies can be readily identified with visual inspection (e.g. Fig. 2I), and reliable estimates can be achieved with a combination of automatic scoring (individual median or optimal threshold) followed by human supervision. However, given the frequency of discrepancies in the automatic scoring of CCRT, additional time investment may be required when analyzing large datasets, and this method may be less suitable for unsupervised analyses. A thorough evaluation of sensitivity is beyond the scope of the current article, but our analysis tool can be tuned to capture, or exclude, subtle movements when necessary by calibrating the detection sensitivity with the video transformation parameters.

Application of automated scoring for thermal limits

Our results indicate that the automated methods individual median and optimal threshold for scoring thermal limits can provide comparable values to manual scoring of CT_{max}, CT_{min}, HDKT and CCRT. The values obtained in this study are also consistent with previous work on thermal limits in *D. melanogaster.* While previous work indicated discrepancies in thermal limit estimates between air- and water-cooled apparatuses (MacLean et al., 2022), CT_{min} and CT_{max} estimates obtained here were within ∼1°C of those from a previous study in our lab using a water-jacketed cylinder (Lecheta et al. (2020), although different lines and rearing conditions were used between the two studies. To further demonstrate the utility of this method beyond *Drosophila*, we applied our automated scoring of CT_{max} to six other insect species (Table S2). The automated methodology scored CT_{max} in every case, despite differences in body structure, size or locomotion. For cases where previous CT_{max} estimates exist for the additional species, the estimate of CT_{max} was in line with expected values, although methodological differences (genetic background, rearing conditions, assay conditions, etc.) make it challenging to directly compare our results with the literature. The CT_{max} estimate for bedbugs (*Cimex lectilarius*) was identical to that previously reported in DeVries et al. (2016), while for the subterranean termite *Reticulotermes flavipes*, which had the largest discrepancy, our CT_{max} estimate was still within 3°C of a previous study (Sponsler and Appel, 1991). When comparing the automated results with visual estimates, where there were differences, the automated method often underestimated the activity endpoint time, likely because of an inability to accurately adjust the 'threshold' of biological activity for every single individual. Thus, while automated measurements of thermal tolerance may slightly vary from classic visual estimates, we see an opportunity for reproducible and high throughput methodologies such as DIME and other similar approaches (e.g. Awde et al., 2020; Laursen et al., 2021; MacLean et al., 2022) to increase sample sizes and standardize endpoints used for insect thermal limits.

Acknowledgements
We acknowledge Yuta Kawarasaki and Kaitlin Donlon for suggestions in the design of arenas, Ioulia Bespalova for feedback on the functionality of previous versions, Luis E. Castañeda and Juan Soto Hernández for suggestions on the implementation of the application DIME, Clare Rittschof for comments on an earlier version of the manuscript, as well as Zach DeVries, Angela J. Sierras, Reddy Palli, Jeffrey L. Howell, Syed Zainulabeuddin, Kenneth O'Dell and Joe Zhou for providing insect samples. We also acknowledge the facility resources provided by the Department of Entomology at the University of Kentucky. Results and some sections of the discussion in this paper are reproduced from the PhD thesis of F.R.P.-G. (Perez-Galvez, 2023).

Competing interests
The authors declare no competing or financial interests.

Author contributions
Conceptualization: F.R.P.-G., D.N.A., N.M.T.; Methodology: F.R.P.-G., S.Z., A.C.W., C.L.C.; Software: F.R.P.-G.; Resources: N.M.T.; Data curation: S.Z., A.C.W.; Writing - original draft: F.R.P.-G.; Writing - review & editing: N.M.T.; Supervision: N.M.T.; Project administration: F.R.P.-G.

Funding
This work was supported by Biotechnology Risk Assessment Grants Program grant 2017-33522-27068 from the USDA National Institute of Food and Agriculture, Hatch Project 1010996 from the USDA National Institute of Food and Agriculture, and National Science Foundation grant OIA-1826689 to N.M.T. D.N.A. was supported by the Ministry of Education, Youth and Sports of the Czech Republic at the time of manuscript submission (project number CZ.02.2.69/0.0/0.0/18_053/0016979).

Data availability
The command-line application DIME can be downloaded from GitHub (https://github.com/fernan9/DIME) and data from Dryad (Perez et al., 2023; https://doi.org/10.5061/dryad.cfxpnvxc2).

References

Addo-Bediako, A., Chown, S. L. and Gaston, K. J. (2000). Thermal tolerance, climatic variability and latitude. *Proc. R. Soc. Lond. Ser. B: Biol. Sci.* **267**, 739-745. doi:10.1098/rspb.2000.1065

Awde, D. N., Fowler, T. E., Pérez-Gálvez, F., Garcia, M. J. and Teets, N. M. (2020). High-throughput assays of critical thermal limits in insects. *J. Vis. Exp.* **160**, e61186. doi:10.3791/61186

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models Usinglme4. *J. Stat. Softw.* **67**, 1-48. doi:10.18637/jss.v067.i01

Bradski, G. (2000). The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* **25**, 120-123. https://www.drdobbs.com/open-source/the-opencv-library/184404319;

Castaneda, L. E., Calabria, G., Betancourt, L. A., Rezende, E. L. and Santos, M. (2012). Measurement error in heat tolerance assays. *J. Therm. Biol.* **37**, 432-437. doi:10.1016/j.jtherbio.2012.03.005

Dallas, H. F. and Rivers-Moore, N. A. (2012). Critical thermal maxima of aquatic macroinvertebrates: towards identifying bioindicators of thermal alteration. *Hydrobiologia* **679**, 61-76. doi:10.1007/s10750-011-0856-4

David, R. J., Gibert, P., Pla, E., Petavy, G., Karan, D. and Moreteau, B. (1998). Cold stress tolerance in *Drosophila*: analysis of chill coma recovery in *D. melanogaster*. *J. Therm. Biol.* **23**, 291-299. doi:10.1016/S0306-4565(98)00020-5

Devries, Z. C., Kells, S. A. and Appel, A. G. (2016). Estimating the critical thermal maximum (CTmax) of bed bugs, Cimex lectularius: comparing thermolimit respirometry with traditional visual methods. *Comp. Biochem. Physiol. A: Mol. Integr. Physiol.* **197**, 52-57. doi:10.1016/j.cbpa.2016.03.003

Garcia-Robledo, C., Kuprewicz, E. K., Staines, C. L., Erwin, T. L. and Kress, W. J. (2016). Limited tolerance by insects to high temperatures across tropical elevational gradients and the implications of global warming for extinction. *Proc. Natl. Acad. Sci. USA* **113**, 680-685. doi:10.1073/pnas.1507681113

Gerken, A. R., Eller, O. C., Hahn, D. A. and Morgan, T. J. (2015). Constraints, independence, and evolution of thermal plasticity: probing genetic architecture of long- and short-term thermal acclimation. *Proc. Natl. Acad. Sci. USA* **112**, 4399-4404. doi:10.1073/pnas.1503456112

Hazell, S. P. and Bale, J. S. (2011). Low temperature thresholds: are chill coma and CTmin synonymous? *J. Insect Physiol.* **57**, 1085-1089. doi:10.1016/j.jinsphys.2011.04.004

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57**, 1-17. doi:10.2307/2334932

Jørgensen, L. B., Malte, H. and Overgaard, J. (2019). How to assess *Drosophila* heat tolerance: unifying static and dynamic tolerance assays to predict heat distribution limits. *Funct. Ecol.* **33**, 629-642. doi:10.1111/1365-2435.13279

Kellermann, V., Loeschcke, V., Hoffmann, A. A., Kristensen, T. N., Fløjgaard, C., David, J. R., Svenning, J. C. and Overgaard, J. (2012). Phylogenetic constraints in key functional traits behind species' climate niches: patterns of desiccation and cold resistance across 95 *Drosophila* species. *Evolution* **66**, 3377-3389. doi:10.1111/j.1558-5646.2012.01685.x

Killick, R. and Eckley, I. (2014). changepoint: an R package for changepoint analysis. *J. Stat. Softw.* **58**, 1-19. doi:10.18637/jss.v058.i03

Kingsolver, J. G. and Umbanhowar, J. (2018). The analysis and interpretation of critical temperatures. *J. Exp. Biol.* **221**, jeb167858. doi:10.1242/jeb.167858

Laursen, S. F., Hansen, L. S., Bahrndorff, S., Nielsen, H. M., Noer, N. K., Renault, D., Sahana, G., Sørensen, J. G. and Kristensen, T. N. (2021). Contrasting manual and automated assessment of thermal stress responses and larval body size in black soldier flies and houseflies. *Insects* **12**, 380. doi:10.3390/insects12050380

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255-268. doi:10.2307/2532051

Lecheta, M. C., Awde, D. N., O'leary, T. S., Unfried, L. N., Jacobs, N. A., Whitlock, M. H., Mccabe, E., Powers, B., Bora, K. and Waters, J. S. (2020). Integrating GWAS and transcriptomics to identify the molecular underpinnings of thermal stress responses in *Drosophila melanogaster*. *Front. Genet.* **11**, 658. doi:10.3389/fgene.2020.00658

Lutterschmidt, W. I. and Hutchison, V. H. (1997a). The critical thermal maximum: data to support the onset of spasms as the definitive end point. *Can. J. Zool.* **75**, 1553-1560. doi:10.1139/z97-782

Lutterschmidt, W. I. and Hutchison, V. H. (1997b). The critical thermal maximum: history and critique. *Can. J. Zool.* **75**, 1561-1574. doi:10.1139/z97-783

Maclean, H. J., Sørensen, J. G., Kristensen, T. N., Loeschcke, V., Beedholm, K., Kellermann, V. and Overgaard, J. (2019). Evolution and plasticity of thermal performance: an analysis of variation in thermal tolerance and fitness in 22 *Drosophila* species. *Philos. Trans. R. Soc. B* **374**, 20180548. doi:10.1098/rstb.2018.0548

Maclean, H. J., Hansen, J. H. and Sørensen, J. G. (2022). Validating the automation of different measures of high temperature tolerance of small terrestrial insects. *J. Insect Physiol.* **137**, 104362. doi:10.1016/j.jinsphys.2022.104362

Martin Bland, J. and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**, 307-310. doi:10.1016/S0140-6736(86)90837-8

Ørsted, M., Rohde, P. D., Hoffmann, A. A., Sørensen, P. and Kristensen, T. N. (2018). Environmental variation partitioned into separate heritable components. *Evolution* **72**, 136-152. doi:10.1111/evo.13391

Perez, F. et al. (2023). Data from: Scoring thermal limits in small insects using open-source, computer assisted motion detection [Dataset]. *Dryad.* doi:10.5061/dryad.cfxpnvxc2

Perez-Galvez, F. R. (2023). Ecological risk assessment of transgenic conditional lethality systems for genetic biocontrol strategies. *PhD thesis*, University of Kentucky.

Rezende, E. L., Castañeda, L. E. and Santos, M. (2014). Tolerance landscapes in thermal ecology. *Funct. Ecol.* **28**, 799-809. doi:10.1111/1365-2435.12268

Sinclair, B. J., Alvarado, L. E. C. and Ferguson, L. V. (2015). An invitation to measure insect cold tolerance: methods, approaches, and workflow. *J. Therm. Biol.* **53**, 180-197. doi:10.1016/j.jtherbio.2015.11.003

Sponsler, R. and Appel, A. (1991). Temperature tolerances of the Formosan and eastern subterranean termites (Isoptera: Rhinotermitidae). *J. Therm. Biol.* **16**, 41-44. doi:10.1016/0306-4565(91)90050-C

Sunday, J. M., Bates, A. E. and Dulvy, N. K. (2011). Global analysis of thermal tolerance and latitude in ectotherms. *Proc. R. Soc. B* **278**, 1823-1830. doi:10.1098/rspb.2010.1295