PROCEEDINGS B

Research

CrossMark
click for updates

Author for correspondence:
Christopher H. Taylor
e-mail: christopher.taylor@nottingham.ac.uk

THE ROYAL SOCIETY
PUBLISHING

Why many Batesian mimics are inaccurate: evidence from hoverfly colour patterns

Christopher H. Taylor, Tom Reader and Francis Gilbert

School of Life Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK

(iD) CHT, 0000-0003-4299-7104

Mimicry is considered a classic example of the elaborate adaptations that natural selection can produce, yet often similarity between Batesian (harmless) mimics and their unpalatable models is far from perfect. Variation in mimetic accuracy is a puzzle, as natural selection should favour mimics that are hardest to distinguish from their models. Numerous hypotheses exist to explain the persistence of inaccurate mimics, but most have rarely or never been tested against empirical observations from wild populations. One reason for this is the difficulty in measuring pattern similarity, a key aspect of mimicry. Here, we use a recently developed method, based on the distance transform of binary images, to quantify pattern similarity both within and among species for a group of hoverflies and their hymenopteran models. This allowed us to test three key hypotheses regarding inaccurate mimicry. Firstly, we tested the prediction that selection should be more relaxed in less accurate mimics, but found that levels of phenotypic variation are similar across most hoverfly species. Secondly, we found no evidence that mimics have to compromise between accuracy to multiple model species. However, we did find that darker-coloured hoverflies are less accurate mimics, which could lead to a trade-off between mimicry and thermoregulation in temperate regions. Our results shed light on a classic problem concerning the limitations of natural selection.

1. Introduction

Charles Darwin regarded mimicry as a beautiful example of the extreme results of natural selection ([1], p. 392), and the topic has since been well studied as a powerful and conspicuous demonstration of the evolution of phenotypes [2]. Batesian mimics are harmless organisms that resemble a more dangerous 'model' in order to deceive potential predators [3], and while some show an astonishing level of similarity to their models, others bear only a passing resemblance. Both theory [4] and experiments [5–7] show that, in practical terms, mimicry is a continuum rather than a simple binary category: inaccurate mimics are attacked less frequently than non-mimics, but more often than more accurate ones (but see [8,9]). We would, therefore, expect the most accurate mimics in a population to have the highest fitness, and that natural selection should drive ever-increasing perfection in resemblance to the model. Contrary to this prediction, there are many examples, including some snakes [10], spiders [11], and hoverflies [12], that seem far from accurate in their mimicry. By exploring this discrepancy between expectation and observation, the study of inaccurate Batesian mimicry provides an excellent opportunity to develop a better understanding of the ecological forces which determine the evolution of phenotypes.

There is no shortage of hypotheses proposed to address the existence of inaccurate mimicry, and these have been well reviewed elsewhere [2,13–15]. Here, we test some of the key hypotheses using hoverflies (Diptera: Syrphidae) as our study organisms, but the hypotheses are equally relevant to other groups of mimics. Hoverflies have been a major focus for studies of inaccurate mimicry, as the taxon comprises a large number of species, many of which are abundant

and widespread, ranging from non-mimetic to highly accurate mimics of various hymenopteran models, with a wide range of inaccurate mimics in between [12,15]. Hoverflies overlap their models extensively in space (with models such as *Apis mellifera* and *Vespula vulgaris* being widespread in the Palearctic), and also in time. Most species of hoverfly first emerge between March and May and remain active until at least September [16], with workers of social Hymenoptera generally reaching peak abundance in July/August [17].

Theoretical explanations for inaccurate Batesian mimicry have produced a number of testable predictions about variation within and among mimetic species. An important group of predictions centre on the cognition and behaviour of the predator, which can be modelled using signal detection theory [4]. This assumes that predators receive information from signals subject to noise, and therefore, uncertainty. Signal detection theory suggests that, past a certain minimum level of similarity, further improvements in mimetic accuracy provide very little decrease in predation risk [18]. Mimics that have reached this critical level of similarity will, therefore, experience relaxed selection. From this, Holloway *et al.* [19] make the prediction that more accurate mimics should show greater phenotypic variation. They suggest that less accurate mimics are under strong selection but lack the genetic variation to evolve closer similarity to the model, and hence have low phenotypic variation.

However, alternative predictions arise if we consider that mimic species may not all be equally attractive to predators. The threshold similarity level described above, beyond which selection is relaxed [18], depends on what has been described as the 'incentive to attack' [20]. A predator is less likely to risk an attack with an uncertain outcome if the cost of attacking a model is high relative to the benefit of consuming a mimic, or if the abundance of models is high relative to the mimics. One possible cause of low incentive to attack is given by Penney *et al.* [21], who argue that smaller mimics have a lower calorific value to the predator, resulting in a low incentive to attack, and hence favouring relatively imperfect mimicry in smaller species. Regardless of the exact reasons behind the costs and benefits to a predator, if a certain group of mimics offer a low incentive to attack, they are predicted to be under relatively relaxed selection by predators compared with other species, and may therefore show greater phenotypic variability.

We must also consider that predators may be influenced by more than one model phenotype. Mathematical models predict that mimics with an intermediate similarity to several model species can be better protected than an accurate mimic of a single model species [14,18], and thus increasing similarity to one model might come at the cost of lower accuracy to another. It is highly likely that predators will encounter more than one model species in their foraging, but the extent to which this influences inaccurate mimicry is not known [14,15].

Finally, if selective pressures other than those imposed by predators influence the mimic's appearance, then inaccurate mimics could represent a trade-off between such opposing pressures. For example, increasing similarity to the model may come with a physiological cost, such as reduced ability to regulate temperature. Hoverfly colour patterns are known to vary with temperature both seasonally and geographically [22], and this variation is thought to confer a survival advantage in response to differing thermoregulatory constraints [23]. In temperate climates, darker coloured insects are able to warm up more quickly [24,25], and thus improve performance in areas such as flight activity [26]. It is highly plausible that such a mechanism underlies colour variation in hoverflies. However, to our knowledge, the effect of this variation on mimetic accuracy has never been assessed. We would expect to see a conflict in temperate regions between the bright colours required for mimicry and dark colours that allow effective temperature regulation.

Among the wealth of theories which seek to explain inaccurate mimicry, most have been studied through mathematical modelling or abstract experiments [2,13]. Only recently has attention turned to a broader perspective of testing the various hypotheses against each other in real systems, which is the only way in which the relative merits of the different hypotheses can be accurately assessed. Penney *et al.* [21] carried out a comparative study of 38 hoverfly species, along with 10 putative models, using both morphological data and human judgement to measure degree of similarity. They found evidence that inaccurate mimics are not just artefacts of human perception, and suggested that no species are intermediate between several models. However, they found a positive relationship between size and mimetic accuracy, which they interpret as evidence for the relaxed selection theory, suggesting that larger hoverflies are more valuable prey and, therefore, under stronger selective pressure.

Another comparative study by Holloway *et al.* [19] investigated the levels of phenotypic variation in a number of hoverfly and wasp species. They used rankings of mimetic accuracy as calculated from behavioural responses of pigeons recorded in Dittrich *et al.* [6], and were consequently limited to the few species used in the pigeon study. Holloway *et al.* [19] found high levels of variation in many species, giving no indication that a lack of genetic variation was limiting the evolution of accuracy. They did not find a clear trend between mimetic accuracy and phenotypic variation, although particularly high variation in the model species and one accurate mimic, *Temnostoma vespiforme*, led them to conclude that relaxed selection may be acting in those cases.

The few empirical studies which have attempted to test predictions about variation in mimetic accuracy have been constrained by the difficulties of generating effective measures of similarity between mimics and their models. It is possible to use predator behaviour to rank similarity (e.g. [6]), but this approach becomes prohibitively expensive if applied to large numbers of specimens, and so in large-scale studies, a mathematical similarity measure is essential. For example, Holloway *et al.* [19] characterized mimic phenotype simply using the proportion of yellow versus black on two tergites of the abdomen. The descriptors that Penney *et al.* [21] used to create a multivariate measure of mimetic accuracy included morphometric data (e.g. antenna length, thorax width, wing length) as well as some summary variables relating to the abdominal pattern (e.g. mean red–green–blue values, number of stripes), but very little about the pattern itself.

Recently, we have developed a new objective measure of mimetic accuracy by comparing entire abdominal patterns using the distance transform method [27]. This method is not intended as a faithful representation of a potential predator's cognitive processes, which in any case are not currently known, but as an objective means of capturing detailed information about pattern variation, beyond simple summary measures such as colour proportions. Nonetheless, our method provides a measure of mimetic accuracy much closer to human and avian estimates than previous empirical measures, even

without the inclusion of any morphometric data [27]. In this study, we use this new methodology to characterize the mimetic patterns of hoverflies in detail, and to test some of the predictions which have emerged from theoretical work. We plot a large number of model and mimic individuals in 'similarity space', giving a picture not only of how species compare with one another in appearance, but also of the variation within species. We then test four predictions associated with three theoretical explanations for the existence of inaccurate mimicry:

1. Relaxed selection
   (a) *Lack of genetic variation*. Less accurate mimics are under strong selection but lack the genetic variation to evolve increased accuracy; more accurate mimic species experience relaxed selection and thus have higher levels of phenotypic variation.
   (b) *Incentive to attack*. Less accurate mimic species have higher levels of phenotypic variation as they provide a lower incentive to attack and are under more relaxed selection.
2. Multiple models. Increasing accuracy to one model decreases accuracy to others; inaccurate mimics represent a compromise between two or more model phenotypes.
3. Thermoregulation. Less accurate mimics have more black in their pattern and hence will be better able to regulate their temperature; there is a trade-off between accurate mimicry and effective thermoregulation.

2. Material and methods

Image processing and dissimilarity calculations were carried out in MATLAB [28]. Statistical analyses were carried out in R v. 3.0.3 [29].

(a) Specimens

Insects were collected using a hand net from wild communities in Nottinghamshire, UK (particularly the Attenborough Nature Reserve) and surrounding areas, during May to October in the years 2012–2014. See electronic supplementary material, table S1 for full details of sampling sites. Target insects were any hoverflies or stinging Hymenoptera bearing a two-colour pattern (usually black and yellow; see example images in electronic supplementary material, figure S1), but excluding bumblebees and their putative mimics, which are notably much hairier than the other taxa encountered (making automated characterization of the abdominal pattern difficult), and which are very likely part of a different mimicry ring [15]. We follow other studies such as Penney *et al.* [21] in excluding male Hymenoptera from the analysis as, not having a sting, their status as models is debatable (they may still be unpalatable to predators due to other factors [5]). Males are also of much lower abundance than females for most of the year, and thus only five specimens were excluded from this study. A total of 954 individuals were identified to species level and sexed using relevant keys [16,17,30].

Specimens were euthanized by freezing, and their legs and wings pinned out to the sides when necessary to give a clear view of the abdomen. They were then placed inside a home-made 'photo studio'—a white $30 \times 18 \times 10$ cm open topped box. A 5 mm scale bar was placed near to the insect. Specimens were photographed from above with a Canon 600D DSLR camera and Tamron 90 mm macro lens under natural outdoor light conditions, in the shade. This method resulted in images that were evenly lit and free from strong reflections or glare. While natural weather variation did lead to some changes in brightness from image to image, this did not affect the analysis because patterns were converted to binary form before comparison (see 'image processing').

(b) Image processing

Images were rotated, cropped, and rescaled to a standard alignment, and an algorithm was applied to remove noise and sharpen edges. An edge detection algorithm was used to find the outline of the abdomen. In some cases, a rough outline was drawn manually and passed to the algorithm as a starting point, to fix cases where the outline was difficult to detect against the background.

The abdomen was automatically segmented into two colour regions (typically black and yellow/orange). Some images (129 out of 954) did not produce clear segmentations, often due to fading of the colours after death (C Taylor 2012, personal observation) and were discarded from further analyses. To quantify the colour proportions in the pattern, we calculated the proportion of pixels within the abdominal image that were classified as 'black' (i.e. the darker of the two segments) after segmentation.

See electronic supplementary material, Text and figure S2, for more detail on the image processing.

(c) Mimetic accuracy

We calculated dissimilarity values for all possible pairings of images within the dataset using the distance transform method [27]. Optimization of the method used translation and scaling in the vertical direction to account for any slight misalignment of the patterns. For some subsequent figures and analyses, it is more intuitive to work with measures of mimetic accuracy than with dissimilarity. To make the conversion, we used the formula $A = 1 - (D/D_{max})$, where A is mimetic accuracy, D is dissimilarity, and D_{max} is the largest dissimilarity value between any two individuals in the overall dataset. This scales mimetic accuracy to run from a minimum of 0 (defined by this particular dataset) to a maximum of 1 (independent of the dataset − identical images). For each individual mimic, we calculated the mean similarity with respect to all individuals of a given model species, to give a measure of mimetic accuracy to that model.

We first tested for sexual dimorphism in the hoverfly species, as males and females may have different levels of mimetic accuracy, or might even resemble different models. For example, it has been suggested that female *Eristalis arbustorum* are bee mimics, while the males mimic wasps [31]. For each mimic species in our dataset for which we have data on at least three males and three females, we tested for dimorphism in both size and pattern. For size, we carried out a Wilcoxon's two-sample test on thorax width data. For pattern, we used distance-based multivariate analysis [32,33] carried out in the program DISTLM5. This allows the equivalent of ANOVA to be carried out directly on distance (dissimilarity) data rather than having to ordinate the data first. Species were considered dimorphic if $p < 0.05$ for either of the above tests, in which case the sexes were treated separately in all subsequent analyses. For species where $p > 0.05$ for both size and pattern, and those with fewer than three individuals in one or other sex, data from males and females were pooled in subsequent analyses. We refer to these groupings as 'species or sex units', or SSUs.

The mimetic accuracy for an SSU was calculated as the mean of the individual values of mimetic accuracy within that SSU, again for each model species separately. We then assigned each SSU a 'best' model, being the potential model for which it has the highest mean accuracy value. Four species of Hymenoptera were treated as the candidate models of the sampled community (electronic supplementary material, figure S1), being the only potential models that were common in our samples ($N > 3$): *V. vulgaris* (common wasp), *V. germanica* (German wasp), *V. crabro* (hornet), and *A. mellifera* (honeybee). We know from both theory [34] and experiments [35] that a model's importance in shaping predator behaviour increases with its abundance, and therefore, we have excluded eight low-abundance ($N \leq 3$) model species from the main analysis. However, we did also repeat our analysis including these rarer model species (see electronic supplementary material, Text).

(i) Relaxed selection

To quantify variation within SSUs, we first ordinated the dissimilarity data by using Principal Coordinates Analysis (PCoA) [36]. We chose this method of ordination, as opposed to non-metric multidimensional scaling, as we considered it important to use a method in which the resulting inter-point distances would be linearly related to the original distance matrix. We did this in order to preserve the magnitude of the variation in the dataset, despite the fact that PCoA assumes that distances between individuals are metric (that is, they obey the triangle inequality), which is not always the case when using distances generated by the distance transform method [27].

On the basis of a scree plot, we chose the first four dimensions of the PCoA as the best representation of the data. Using these four dimensions, we calculated the centroid for each SSU, and then the distance, z, of each individual to its corresponding centroid. The mean of a group's z values provides a measure of within-group variability [33].

When testing the relationship between mimetic accuracy and within-taxon variability in accuracy, using raw similarity values as a measure of mimetic accuracy is not appropriate. If a model and mimic species overlap in phenotypic space, we risk creating a circular argument. Mimics that are more variable will inevitably show lower accuracy, as a greater spread in phenotypic space will lead to larger distances (on average) to the model phenotype. For this test, therefore, we used a different measure of mimetic accuracy that is not affected by the phenotypic variability. After ordination using PCoA, we calculated centroid points for mimic and model species and defined (in)accuracy as the distance from a mimic's centroid to the closest model centroid.

To test for an influence of mimetic accuracy on within-taxon variability, we ran a generalized least-squares model (GLS) [37] in the R package 'ape' v. 3.1-1 [38]. GLS is equivalent to a general linear model, but with the inclusion of a correlation matrix derived from the species' phylogeny to control for relatedness among species. We used mean z value for an SSU as the response and mean mimetic accuracy and mean thorax width (plus their interaction) as predictors. Thorax width was included in the model as a proxy for size [39], because Penney et al. [21] argued that larger hoverflies should offer a larger 'incentive to attack' due to their greater nutritional value. The width of the thorax at the base of the wings was measured in IMAGEJ [40] using the unprocessed images, using the 5 mm bar in each image to set the scale. Note that in the early stages of the project, photographs did not include a scale bar, and therefore, in some cases (e.g. electronic supplementary material, table S2) samples for size measurements are smaller than for other measures such as pattern.

We tested the model under two different evolutionary scenarios: Brownian motion (BM) evolution and Ornstein–Uhlenbeck (OU) evolution (similar to BM, but traits are constrained towards an 'optimum' value). These different scenarios were represented by two different correlation matrices passed to the GLS model, calculated from a composite phylogeny (electronic supplementary material, figure S3) based on information from Rotheray and Gilbert [41], and Ståhls et al. [42]. For both females and males, the OU evolutionary model was found to be a significantly better fit to the data (females: likelihood ratio (LR) = 11.71, $p = 0.0006$; males: LR = 6.10, $p = 0.014$; both d.f. = 1) and was used for subsequent analysis. We then used backwards stepwise model simplification with likelihood ratio tests to find the minimum adequate model. In order to allow for sexual dimorphism, we conducted two separate analyses, one with data from only female individuals and the other with only males.

(ii) Multiple models

To test for a potential trade-off in similarity to multiple models, we tested within SSUs for correlation (using Pearson's r) in mimetic accuracy towards the four main model species. A negative correlation would imply that, for a given SSU, increasing similarity to one model comes at the cost of decreased similarity to another. We tested all SSUs for which we had data on at least six individuals.

(iii) Thermoregulation

We tested for a trade-off between accuracy and the extent of black in the pattern (proportion black) using a Markov Chain Monte Carlo generalized linear mixed model, implemented in the R package 'MCMCglmm' [43]. Again, this method allowed us to control for phylogenetic relatedness among species. Accuracy of individual mimics to their closest model was the response variable, logit transformed for normality of residuals. Fixed effects were the proportion black, thorax width, sex, and season, along with all two-way interactions. Thorax width was included as a proxy for size (see above), which can have a major impact on thermoregulation [44]. Season was included because selection on thermoregulation may vary according to the time of year. We categorized season as 'early' (to 8 August) or 'late' (9 August onwards) splitting at the date of the median sample, which also fell roughly halfway between the first and last sampling days. We also conducted a more complex analysis in which time of year was treated as a continuous variable, including a quadratic term, which gave very similar results. Species was included as a random effect, and we calculated a covariance structure for the random effect based on the phylogenetic tree (electronic supplementary material, figure S3; also see 'relaxed selection' above). We used backwards stepwise model simplification based on p-values to find the minimum adequate model. Note that in figure 2, proportion black and thorax width are binned for ease of interpretation, but they were treated as continuous in the analysis.

3. Results

We examined pattern similarity among 697 hoverfly (54 species) and 128 hymenopteran individuals (12 species). We found evidence for size dimorphism in seven of the mimic species in our dataset, and for pattern dimorphism in a further 11 (electronic supplementary material, table S2), giving a total of 72 SSUs. Compared against the four most abundant species of Hymenoptera from our samples, 51 SSUs were classed as mimics of V. vulgaris, 11 of A. mellifera, 7 of Vespa crabro, and 3 of V. germanica. The level of mimetic accuracy to the assigned model varied from 0.55 to 0.87 (electronic supplementary material, table S3, figure S4).

(a) Relaxed selection

If inaccurate mimics have insufficient genetic variation to reach a level of protection at which selection becomes relaxed, we predict a positive correlation between pattern variability within species and similarity to the model. Alternatively, if less accurate mimic species provide a low incentive for predators to attack, for example because of a low calorific value, we predict a negative correlation. However, after controlling for shared ancestry, phenotypic variability was not significantly associated with either mimetic accuracy or body size (thorax width) in either males or females (table 1 and figure 1; see also electronic supplementary material, table S4).

(b) Multiple models

If mimetic accuracy is limited by a trade-off among similarities to several models, we predict that similarity to different model species should be negatively correlated. However, almost all

Table 1. GLS models of within-species variability. The contribution of each predictor to the model was assessed using a likelihood ratio test. All tests had Δd.f. = 1. Sample size was 32 for females and 34 for males.

predictor	likelihood ratio	*p*-value
female		
accuracy : size	0.1	0.748
accuracy	0.82	0.365
size	1.09	0.296
male		
accuracy : size	0.63	0.427
accuracy	0.87	0.350
size	0.73	0.392

Figure 1. The relationship between pattern variability (mean *z* value) of an SSU and its mimetic accuracy.

SSUs show either a significant positive correlation or no significant correlation among similarity values to the four main model species (electronic supplementary material, table S5). There was only one negative correlation with $p < 0.05$: in males of *Syrphus ribesii*, accuracy to *A. mellifera* was negatively correlated with accuracy to *Vespa crabro* ($r = -0.56$, $p = 0.009$, $N = 21$). Under the null hypothesis, if all tests were independent, we would expect 10 negative correlations through type I error on average.

(c) Thermoregulation

If mimetic accuracy is limited by a trade-off with thermoregulation, we predict a negative correlation between similarity to the model and the proportion of the pattern that is black. Having controlled for shared ancestry, there is a significant negative interaction between proportion black and thorax width ($p = 0.040$; table 2). When combined with the other estimated coefficients (table 2) this indicates that those mimics with a greater proportion of black on their abdomen tend to be less accurate to their model, and that this trend is particularly strong in larger mimics (figure 2). There is a significant effect of sex, with females in general being more accurate ($p < 0.001$). In addition, both proportion black ($p < 0.001$) and thorax width ($p < 0.001$) interact with sex, with females showing a weaker version of the trend described above. These trends observed in colour, size, and sex are evident even having accounted for seasonal differences in mimetic accuracy (table 2 and figure 2; see also electronic supplementary material, tables S6–S7 and figure S5).

4. Discussion

By comparing colour patterns using the distance transform method [27], we have been able to quantify in detail the mimetic relationships in a community of insects, including variation both within and among species. The lack of a trend between accuracy and phenotypic variation suggests that inaccurate mimics are not accounted for by the fact that they have not been able to evolve to the point of maximum protection (Prediction 1a) or by relaxed selection caused by a reduced incentive of predators to attack (Prediction 1b). Rather, the data suggest that inaccurate phenotypes represent the result of a trade-off between opposing selective pressures. A trade-off caused by selection for similarity to multiple models
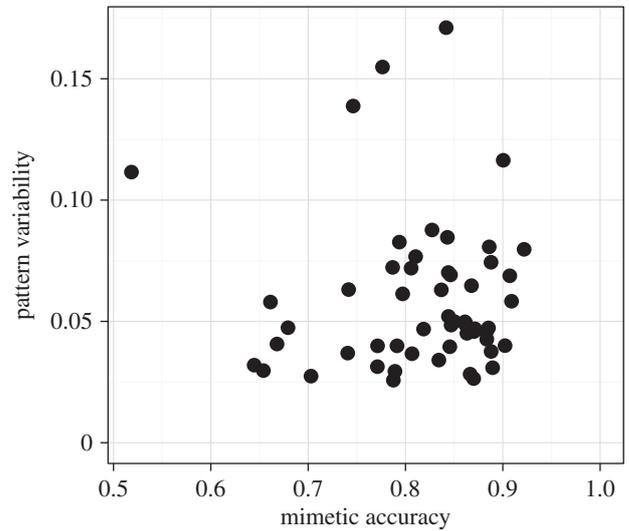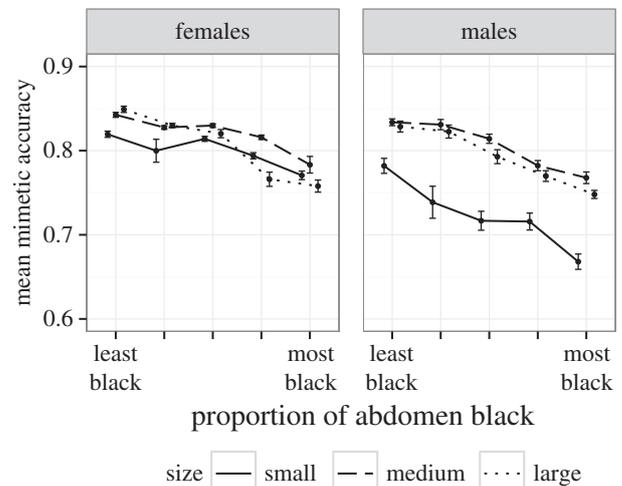

Figure 2. The effect of colour ratio on mimetic accuracy. Hoverfly individuals have been binned into three size categories in equal proportions: small (thorax up to 2.5 mm wide; solid line), medium (2.6 to 3.8 mm; dashed line), and large (3.9 mm or more; dotted line), and five colour categories (up to 52% black, 53–59% black, 60–66% black, 67–74% black, and 75% or more black). Error bars show ± standard error (s.e.). Note truncation of the *y*-axis.

(Prediction 2) is not supported, but the results point towards a hitherto unexplored role for thermoregulation in limiting the adaptive value of increased accuracy (Prediction 3).

The absence of a trend in phenotypic variation with mimetic accuracy and the relatively high levels of phenotypic variation are broadly in line with the results from Holloway *et al.* [19]. Therefore, it seems unlikely that inaccurate mimics are limited by a lack of genetic variation. We cannot tell from these data how much of the variation is heritable; at least some will be attributable to measurement error, and some to phenotypic plasticity, as (for example) adult patterns are known to change with the temperature experienced by the puparium [45]. However, the few studies of the genetic component of pattern variation in hoverfly species found a high level of heritability in those cases [46,47].

The relaxed selection hypothesis predicts that, above a certain level of similarity, any further improvements in mimetic

Table 2. MCMCglmm model of mimetic accuracy. Accuracy was logit transformed for normality. SSU was included as a random effect, with a variance structure that accounts for phylogenetic relatedness. Backwards model selection was used on the basis of the p-values. Posterior means are quoted for coefficients of all predictors present in the minimum adequate model. All factors have d.f. $= 1$, $N = 638$.

predictor	posterior mean	pMCMC
intercept	1.34	<0.001
proportion black	0.158	0.614
thorax width	0.052	0.434
sex (F)	0.426	<0.001
season (late)	−0.090	0.066
proportion black: thorax width	−0.204	0.040
proportion black: sex (F)	0.396	<0.001
thorax width: sex (F)	−0.188	<0.001
sex (F): season (late)	0.053	0.030
thorax width: season (late)	0.045	<0.001
proportion black: season (late)		0.104

accuracy are selectively neutral [18]. Penney et al. [21] found a correlation between size and morphometric similarity to the model, and argued that smaller prey items are less valuable, and so relaxed selection allows the persistence of inaccurate mimicry in smaller hoverflies. However, a predator's optimal diet depends not only on the calorific value of the prey but also on search and handling times [48], and it is not clear whether large hoverflies provide the best trade-off in that regard. Furthermore, although Penney et al. [21] found that larger hoverflies tend to be more similar to their models in terms of morphology, our results reveal a more complicated relationship between pattern similarity and size. There is no direct effect of size on accuracy (table 2) although there is an interaction with the colour proportions of the abdomen (see below), and in the case of males, the smallest are indeed the least accurate (figure 2). Most importantly though, our data show no association between phenotypic variation and either size or mimetic accuracy. While our results do not rule out the possibility that selection on mimicry in hoverflies may be relaxed, they do show that relaxed selection is not connected with a species' level of mimetic accuracy or its size, and thus cannot provide an explanation for the observed variation in mimetic accuracy.

We are left with the likely explanation that there is some kind of opposing selective pressure that is balanced against the advantage of increased mimetic accuracy. The multiple models hypothesis provides one possibility. In terms of shape, hoverflies are clearly distinct from Hymenoptera, with none occupying phenotypes intermediate to two or more model species [21]. In terms of pattern, the distinction is less clear. After ordination in two-dimensional space, there are a large number of hoverfly individuals that, for example, occupy the space in between A. mellifera and Vespula spp. (electronic supplementary material, figure S4), but distinguishing an adaptive explanation from random placement is difficult. Crucially, for each species of mimic, there is either no correlation or a positive correlation among similarity values to each potential model species. This implies that, at least in

terms of pattern, there is no multi-model trade-off: assuming the observed variation has an underlying genetic component, it would be possible for each mimic to improve its similarity to one or more models without compromising similarity to others. We cannot rule out multiple models having an influence on the phenotype of a mimic, but we can conclude that the multiple models hypothesis is not sufficient to explain the observed levels of inaccuracy.

By contrast, a trade-off between mimicry and thermoregulation is consistent with our data. Hoverflies maintain a temperature excess (a body temperature above that of the surrounding air) through a combination of basking and shivering [49]. Darker coloured insects absorb more solar radiation, and therefore, can heat up more rapidly [24,25], so we expect darker hoverflies to be at a fitness advantage in cooler conditions. More rapid temperature gain during basking will reduce the opportunity cost of thermoregulation as well as possibly reducing predation risk. In support of this, a number of hoverfly species have been found to show seasonal variation in their colour patterns, with darker morphs being more common outside the summer months [45], which is thought to have an adaptive function in relation to temperature regulation [23].

However, the results of our study show that the thermoregulatory benefits of darker patterns will also likely be associated with a reduction in mimetic accuracy. To be a perfect mimic of V. vulgaris, the most abundant model in our samples would require the amount of black on the abdomen to be limited to 51%, but almost all hoverflies surveyed were above this value (electronic supplementary material, table S3). Aposematic signals are known to constrain temperature regulation, as observed in the moth Parasemia plantagenis [50]. Moths with more black on their body were able to warm up more quickly, but suffered increased predation due to a less effective warning signal. Thus, it is highly plausible that hoverfly colour patterns are constrained by their thermoregulation function. By contrast, wasp abdominal patterns are likely to be less constrained, as they do not rely much on basking for thermoregulation; social wasps achieve a high temperature excess through endothermy before they even leave their nest [51].

Interestingly, we find that the constraint on the colour pattern seems to be stronger in larger individuals, as revealed by the significant interaction between the proportion black and size. Larger insects are able to maintain a higher temperature relative to the ambient, but have slower heating rates [52]. Thus, any differences in rates of warming caused by colour are likely to have a greater effect on fitness in larger than in smaller insects, the latter being unable to depart far from ambient temperature and so rates of warming are less likely to be a relevant factor. Indeed, both theoretical predictions [44] and physical models [53] have shown that colour should have a greater effect on temperature in larger organisms.

We also show that female hoverflies tend to be significantly better mimics than males, suggesting that the evolutionary pressures experienced by the sexes on their appearance are different. A similar observation has been made in butterflies, with females of some species being closer in colour to their models than males [54,55], and many others in which mimicry is entirely restricted to the females [56]. A number of reasons have been suggested to explain those differences, including increased vulnerability of females to predators [57], conflict with intra-sexual signalling in males [58], and facilitating species recognition during mating [55]. These possibilities merit further investigation in hoverflies.

An alternative explanation that is consistent with a trade-off between accuracy and colour ratio could be that darker patterns are more cryptic to predators. It is possible that, as well as affecting mimetic accuracy and thermoregulation, the abdominal colour ratio may also affect the conspicuousness of the pattern. This potential explanation has received little attention in the literature, but it seems likely that, due to their high levels of activity, hoverflies are conspicuous regardless of their exact colour pattern. Even non-mimetic hoverflies are not considered cryptic [12].

For models, our study focused on four common species of Hymenoptera which are often regarded as the targets of mimicry in European hoverfly communities [15], but we caught a number of other hymenopteran species in small numbers, which could potentially also serve as models. The lower abundance and/or visibility of these species during our collection suggests that predators too will encounter them at a low rate, and therefore, their importance as models is likely lower than those species that are widespread and conspicuous [35]. Nonetheless, conclusions are similar when we incorporate these rarer model species into the analysis (see electronic supplementary material, Text). We also note that the four common model species from our study all increase in abundance during late summer/early autumn, and that this change could potentially affect the dynamics of the mimetic community. However, the relationship between colour and mimetic accuracy cannot be explained by seasonal effects, as it was observed even after seasonal variation was taken into account.

The phenotypic correlations we have described are consistent with a trade-off between mimicry and thermoregulation, but we acknowledge that, due to the comparative nature of this study, we have not been able to test this trade-off directly. As we have discussed, the mechanisms that we suggest may be responsible for the observed correlation are consistent with what is known about mechanisms of insect thermoregulation. Further work is now needed to test the effects of colour variation on both predation and temperature of hoverflies in an experimental setting. Comparison of mimetic communities from different climates may also provide a fruitful means of examining the conflict between mimicry and thermoregulation in more detail.

References

1. Darwin F. 1887 *The life and letters of Charles Darwin, vol. 2*. London, UK: John Murray.

2. Ruxton GD, Sherratt TN, Speed MP. 2004 *Avoiding attack: the evolutionary ecology of crypsis, warning signals, and mimicry*. Oxford, UK: Oxford University Press.

3. Bates HW. 1862 XXXII. Contributions to an insect fauna of the Amazon valley. Lepidoptera: Heliconidæ. *Trans. Linn. Soc. Lond.* **23**, 495–566. (doi:10.1111/j.1096-3642.1860.tb00146.x)

4. Oaten A, Pearce CEM, Smyth MEB. 1975 Batesian mimicry and signal detection theory. *Bull. Math. Biol.* **37**, 367–387. (doi:10.1007/BF02459520)

5. Mostler G. 1935 Beobachtungen zur frage der wespenmimikry [Observations on the question of wasp mimicry]. *Zoomorphology* **29**, 381–454. (doi:10.1007/bf00403719)

6. Dittrich W, Gilbert F, Green P, Mcgregor P, Grewcock D. 1993 Imperfect mimicry: a pigeon's perspective. *Proc. R. Soc. Lond. B* **251**, 195–200. (doi:10.1098/rspb.1993.0029)

7. Mappes J, Alatalo RV. 1997 Batesian mimicry and signal accuracy. *Evolution* **51**, 2050–2053. (doi:10.2307/2411028)

8. Valkonen JK, Nokelainen O, Mappes J. 2011 Antipredatory function of head shape for vipers and their mimics. *PLoS ONE* **6**, e22272. (doi:10.1371/journal.pone.0022272)

9. Hossie TJ, Sherratt TN. 2013 Defensive posture and eyespots deter avian predators from attacking caterpillar models. *Anim. Behav.* **86**, 383–389. (doi:10.1016/j.anbehav.2013.05.029)

10. Greene HW, McDiarmid RW. 1981 Coral snake mimicry: does it occur? *Science* **213**, 1207–1212. (doi:10.1126/science.213.4513.1207)

11. Pekár S, Jarab M. 2011 Assessment of color and behavioral resemblance to models by inaccurate myrmecomorphic spiders (Araneae). *Invertebr. Biol.* **130**, 83–90. (doi:10.1111/j.1744-7410.2010.00217.x)

12. Rotheray GF, Gilbert F. 2011 *The natural history of hoverflies*. Cardigan, UK: Forrest Text.

13. Kikuchi DW, Pfennig DW. 2013 Imperfect mimicry and the limits of natural selection. *Q. Rev. Biol.* **88**, 297–315. (doi:10.1086/673758)

14. Edmunds M. 2000 Why are there good and poor mimics? *Biol. J. Linn. Soc.* **70**, 459–466. (doi:10.1111/j.1095-8312.2000.tb01234.x)

15. Gilbert F. 2005 The evolution of imperfect mimicry. In *Insect evolutionary ecology* (eds M Fellowes, G Holloway, J Rolff), pp. 231–288. Wallingford, UK: CABI.

16. Stubbs AE, Falk SJ. 2002 *British hoverflies: an illustrated identification guide*. Reading, UK: British Entomological and Natural History Society.

17. Richards OW. 1980 *Scolioidea, Vespoidea and Sphecoidea; Hymenoptera, Aculeata*. London, UK: Royal Entomological Society of London.

18. Sherratt TN. 2002 The evolution of imperfect mimicry. *Behav. Ecol.* **13**, 821–826. (doi:10.1093/beheco/13.6.821)

19. Holloway G, Gilbert F, Brandt A. 2002 The relationship between mimetic imperfection and phenotypic variation in insect colour patterns. *Proc. R. Soc. Lond. B* **269**, 411–416. (doi:10.1098/rspb.2001.1885)

20. Johnstone RA. 2002 The evolution of inaccurate mimics. *Nature* **418**, 524–526. (doi:10.1038/nature00845)

21. Penney HD, Hassall C, Skevington JH, Abbott KR, Sherratt TN. 2012 A comparative analysis of the evolution of imperfect mimicry. *Nature* **483**, 461–464. (doi:10.1038/nature10961)

22. Holloway GJ. 1993 Phenotypic variation in colour pattern and seasonal plasticity in Eristalis hoverflies (Diptera: Syrphidae). *Ecol. Entomol.* **18**, 209–217. (doi:10.1111/j.1365-2311.1993.tb01092.x)

23. Ottenheim MM, Wertheim B, Holloway GJ, Brakefield PM. 1999 Survival of colour polymorphic Eristalis arbustorum hoverflies in semi field conditions. *Funct. Ecol.* **13**, 72–77. (doi:10.1046/j.1365-2435.1999.00284.x)

24. Kingsolver JG. 1987 Evolution and coadaptation of thermoregulatory behavior and wing pigmentation pattern in pierid butterflies. *Evolution* **41**, 472–490. (doi:10.2307/2409250)

25. Willmer PG, Unwin DM. 1981 Field analyses of insect heat budgets: reflectance, size and heating rates. *Oecologia* **50**, 250–255. (doi:10.1007/BF00348047)

26. Ellers J, Boggs CL. 2004 Functional ecological implications of intraspecific differences in wing

7

rspb.royalsocietypublishing.org   Proc. R. Soc. B 283: 20161585

melanization in *Colias* butterflies. *Biol. J. Linn. Soc.* **82**, 79–87. (doi:10.1111/j.1095-8312.2004.00319.x)

27. Taylor CH, Gilbert F, Reader T. 2013 Distance transform: a tool for the study of animal colour patterns. *Methods Ecol. Evol.* **4**, 771–781. (doi:10.1111/2041-210x.12063)

28. MATLAB. 2012 *MATLAB*. Natick, MA: The Mathworks.

29. R Core Team. 2014 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

30. Perkins RCL. 1919 The British species of *Andrena* and *Nomada*. *Trans. Entomol. Soc. Lond.* **1919**, 218–319.

31. Heal JR. 1981 Colour patterns of Syrphidae. III. Sexual dimorphism in *Eristalis arbustorum*. *Ecol. Entomol.* **6**, 119–127. (doi:10.1111/j.1365-2311.1981.tb00600.x)

32. Anderson MJ. 2001 A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32–46. (doi:10.1111/j.1442-9993.2001.01070.pp.x)

33. McArdle BH, Anderson MJ. 2001 Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297. (doi:10.1890/0012-9658(2001)082[0290:fmmtcd]2.0.co;2)

34. Getty T. 1985 Discriminability and the sigmoid functional response: how optimal foragers could stabilize model-mimic complexes. *Am. Nat.* **125**, 239–256. (doi:10.1086/284339)

35. Lindström L, Alatalo RV, Mappes J. 1997 Imperfect Batesian mimicry—the effects of the frequency and the distastefulness of the model. *Proc. R. Soc. Lond. B* **264**, 149–153. (doi:10.1098/rspb.1997.0022)

36. Legendre P, Legendre L. 1998 *Numerical ecology*, 2nd English edn. Amsterdam, The Netherlands: Elsevier.

37. Grafen A. 1989 The phylogenetic regression. *Phil. Trans. R. Soc. B* **326**, 119–157. (doi:10.1098/rstb.1989.0106)

38. Paradis E, Claude J, Strimmer K. 2004 APE: analysis of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)

39. Gilbert FS. 1985 Morphometric patterns in hoverflies (Diptera, Syrphidae). *Proc. R. Soc. Lond. B* **224**, 79–90. (doi:10.1098/rspb.1985.0022)

40. Abràmoff MD, Magalhães PJ, Ram SJ. 2004 Image processing with ImageJ. *Biophoton. Int.* **11**, 36–42.

41. Rotheray G, Gilbert F. 1999 Phylogeny of Palaearctic Syrphidae (Diptera): evidence from larval stages. *Zool. J. Linn. Soc.* **127**, 1–112. (doi:10.1111/j.1096-3642.1999.tb01305.x)

42. Ståhls G, Hippa H, Rotheray G, Muona J, Gilbert F. 2003 Phylogeny of Syrphidae (Diptera) inferred from combined analysis of molecular and morphological characters. *Syst. Entomol.* **28**, 433–450. (doi:10.1046/j.1365-3113.2003.00225.x)

43. Hadfield JD. 2010 MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22. (doi:10.18637/jss.v033.i02)

44. Stevenson RD. 1985 The relative importance of behavioral and physiological adjustments controlling body temperature in terrestrial ectotherms. *Am. Nat.* **126**, 362–386. (doi:10.2307/2461361)

45. Holloway G, Marriott C, Crocker HJ. 1997 Phenotypic plasticity in hoverflies: the relationship between colour pattern and season in *Episyrphus balteatus* and other Syrphidae. *Ecol. Entomol.* **22**, 425–432. (doi:10.1046/j.1365-2311.1997.00096.x)

46. Conn DLT. 1972 The genetics of the bee-like patterns of *Merodon equestris*. *Heredity* **28**, 379–386. (doi:10.1038/hdy.1972.48)

47. Heal JR. 1979 Colour patterns of Syrphidae I. Genetic variation in the dronefly *Eristalis tenax*. *Heredity* **42**, 223–236. (doi:10.1038/hdy.1979.24)

48. Pyke GH, Pulliam HR, Charnov E. 1977 Optimal foraging: a selective review of theory and tests. *Q. Rev. Biol.* **52**, 137–154. (doi:10.1086/409852)

49. Morgan KR, Heinrich B. 1987 Temperature regulation in bee- and waspmimicking syrphid flies. *J. Exp. Biol.* **133**, 59–71.

50. Hegna RH, Nokelainen O, Hegna JR, Mappes J. 2013 To quiver or to shiver: increased melanization benefits thermoregulation, but reduces warning signal efficacy in the wood tiger moth. *Proc. R. Soc. B* **280**, 20122812. (doi:10.1098/rspb.2012.2812)

51. Heinrich B. 1984 Strategies of thermoregulation and foraging in two vespid wasps, *Dolichovespula maculata* and *Vespula vulgaris*. *J. Comp. Physiol. B* **154**, 175–180. (doi:10.1007/BF00684142)

52. Digby PSB. 1955 Factors affecting the temperature excess of insects in sunshine. *J. Exp. Biol.* **32**, 279–298.

53. Shine R, Kearney M. 2001 Field studies of reptile thermoregulation: how well do physical models predict operative temperatures? *Funct. Ecol.* **15**, 282–288. (doi:10.1046/j.1365-2435.2001.00510.x)

54. Su S, Lim M, Kunte K. 2015 Prey from the eyes of predators: Color discriminability of aposematic and mimetic butterflies from an avian visual perspective. *Evolution* **69**, 2985–2994. (doi:10.1111/evo.12800)

55. Llaurens V, Joron M, Théry M. 2014 Cryptic differences in colour among Müllerian mimics: how can the visual capacities of predators and prey shape the evolution of wing colours? *J. Evol. Biol.* **27**, 531–540. (doi:10.1111/jeb.12317)

56. Kunte K. 2009 Female-limited mimetic polymorphism: a review of theories and a critique of sexual selection as balancing selection. *Anim. Behav.* **78**, 1029–1036. (doi:10.1016/j.anbehav.2009.08.013)

57. Ohsaki N. 1995 Preferential predation of female butterflies and the evolution of Batesian mimicry. *Nature* **378**, 173–175. (doi:10.1038/378173a0)

58. Lederhouse RC, Scriber JM. 1996 Intrasexual selection constrains the evolution of the dorsal color pattern of male Black Swallowtail butterflies, *Papilio polyxenes*. *Evolution* **50**, 717–722. (doi:10.2307/2410844)

8

rspb.royalsocietypublishing.org *Proc. R. Soc. B* **283**: 20161585