

Mapping the evolution of accurate Batesian mimicry of social wasps in hoverflies

Alice Leavey¹, Christopher H. Taylor¹, Matthew R. E. Symonds², Francis Gilbert¹, Tom Reader¹

¹School of Life Sciences, University of Nottingham, Nottingham, NG7 2RD, UK.

²Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Burwood, Victoria 3125, Australia.

Corresponding author: Alice Leavey [AliceLeavey@outlook.com]

Author contributions: T.R. and A.L. conceived and designed the study. T.R. and F.G. supervised the project. A.L. collected the data and performed each analysis. C.H.T. wrote the MATLAB code and assisted image analysis. F.G. provided hoverfly expertise and M.R.E.S. assisted with the phylogenetic comparative analysis. A.L. and T.R. took the lead in writing the manuscript, and all authors provided critical feedback and helped shape the research, analysis and manuscript.

Conflicts of interest: We declare that we have no conflicts of interest.

Acknowledgements: We would like to thank Patrick Reader for coding the website for testing our student volunteers.

Data accessibility statement: Datasets available on Dryad (doi:10.5061/dryad.15dv41nxx).

1 **Abstract**

2 Hoverflies (Diptera: Syrphidae) provide an excellent opportunity to study the
3 evolution of Batesian mimicry, where defenceless prey avoid predation by evolving
4 to resemble defended 'model' species. While some hoverflies beautifully resemble
5 their hymenopteran models, others seem to be poor mimics or are apparently non-
6 mimetic. The reasons for this variation are still enigmatic despite decades of
7 research. Here, we address this issue by mapping social-wasp mimicry across the
8 phylogeny of Holarctic hoverflies. Using the 'distance transform' technique, we
9 calculate an objective measure of the abdominal pattern similarity between 167
10 hoverfly species and a widespread putative model, the social wasp, *Vespula*
11 *germanica*. We find that good wasp mimicry has evolved several times, and may
12 have also been lost, leading to the presence of non-mimics deep within clades of
13 good mimics. Body size was positively correlated with similarity to the model,
14 supporting previous findings that smaller species are often poorer mimics.
15 Additionally, univoltine species were less accurate wasp mimics than multivoltine
16 and bivoltine species. Hence, variation in the accuracy of Batesian mimics may
17 reflect variation in the opportunity for selection caused by differences in prey value
18 or signal perception (influenced by body size) and phenology or generation time
19 (influenced by voltinism).

Keywords: *Batesian mimicry; evolution; Syrphidae; image analysis; similarity;*
distance transform

20 **1 Introduction**

21 Batesian mimicry, where palatable prey avoid predation by evolving features to
22 resemble defended model species (Bates, 1862), not only provides an iconic
23 example of adaptation by natural selection, but also presents a paradox that has
24 challenged evolutionary theory for the last 159 years (Ruxton *et al.*, 2018; Gilbert,
25 2005). Theory predicts that constant selection pressures imposed by predation
26 should improve mimetic accuracy (Dittrich *et al.*, 1993; Cuthill and Bennett, 1993;
27 Edmunds, 2000; Gilbert, 2005, Rotheray and Gilbert, 2011). However, mimicry is
28 frequently far from perfect (Speed and Ruxton, 2010; Edmund and Reader, 2014;
29 Taylor *et al.*, 2016a). Attempts to comprehend the existence of imperfect mimicry
30 have produced an extensive series of hypotheses (see McLean *et al.*, 2019, for
31 a review). While some of these hypotheses are now regarded as implausible,
32 great uncertainty remains over which factors are most important in the
33 persistence of imperfect mimicry.

34 One of the best-known systems for the study of imperfect mimicry is provided by
35 hoverfly mimics (Diptera: Syrphidae), which are probably defenceless, and their
36 harmful hymenopteran models. Many hoverflies imitate Hymenoptera
37 behaviourally (Golding *et al.*, 2005; Penney *et al.*, 2014), acoustically (Moore and
38 Hassall, 2016) and morphologically, in the form of colour, pattern, shape and size
39 (Howarth *et al.*, 2004; Penney *et al.*, 2012; Taylor *et al.*, 2017). However, many
40 supposedly mimetic hoverflies do not accurately resemble their putative models,
41 and others are apparently not mimetic at all. The hoverfly clade therefore provides
42 an ideal opportunity to study how mimetic accuracy has evolved.

43 The study of Batesian mimicry is often hampered by difficulties in defining and
44 quantifying mimicry. Hoverflies have typically been classified as Batesian mimics
45 based on behavioural studies using putative or model predators under controlled
46 conditions, or entirely subjectively, and often inconsistently, by humans (Taylor
47 *et al.*, 2013; Edmunds and Reader, 2014). Even attempts to quantify mimicry
48 more objectively have relied on somewhat *ad hoc* selections of variables or
49 landmarks, often using features which will be perceived very differently
50 depending on the signal receiver (e.g., RGB colour values) (Dittrich *et al.*, 1993;
51 Azmeh *et al.*, 1998; Holloway *et al.*, 2002; Penney *et al.*, 2012). Consequently,
52 our understanding of variation in the accuracy of mimicry among hoverfly species
53 may be at odds with the perception of real predators in the wild. Furthermore, the
54 mimetic status of many hoverflies, especially those that are not conspicuous to
55 the human eye, remains completely unknown.

56 Correlations between mimicry and life-history traits can provide important insights
57 into the factors that have driven the evolution of mimicry. For instance, we might
58 expect mimicry to be related to body size because larger species are more
59 conspicuous to predators, or more valuable prey, while smaller species may
60 benefit more from other anti-predation strategies such as crypsis (Holen and
61 Johnstone, 2004). Wilson *et al.* (2013) found that body size does not correlate
62 strongly with mimetic fidelity in hoverflies, but they did not account for phylogeny
63 (and hence shared evolutionary history) in their analysis. By contrast, a
64 phylogenetically controlled analysis suggested that large hoverfly species are
65 indeed better mimics (Penney *et al.*, 2012). However, neither of these studies
66 explicitly considered hoverflies which are thought to be non-mimics. Studies

67 examining mimicry in coral snakes have found that good Batesian mimicry could
68 gradually evolve from non-mimetic ancestral species, and that maladaptive
69 mimetic patterns can break down, resulting in poor mimics being deeply nested
70 in a clade of good mimics (Kikuchi and Pfennig, 2010; Hodson and Lehtinen,
71 2017). However, life history traits that could be associated with the evolution of
72 mimicry, such as diet or body size, were not considered in these analyses.
73 Additionally, the relative abundance and phenology of mimics and models can
74 impact the selection pressure for good mimicry, factors that are likely to be
75 influenced in insects by voltinism, which can vary substantially among species
76 (Howarth and Edmunds, 2000; Finkbeiner *et al.*, 2018; Hassal *et al.*, 2019). Only
77 by analysing life history traits and phylogenetic history together can we make
78 clear inferences about the evolvability of mimetic accuracy, but this has yet to be
79 attempted for any large taxonomic group, such as the Syrphidae (Gilbert, 2005;
80 Rotheray and Gilbert, 2011).

81 In this study, we build on previous attempts to quantify variability in visual mimetic
82 accuracy among hoverfly species, and to identify the possible drivers of that
83 variability, with a comprehensive phylogenetically-controlled analysis of hoverfly
84 abdominal patterns, features which are detectable by almost any visual system.
85 The key questions we address are: (i) **how has the accuracy of wasp mimicry**
86 **evolved across the hoverfly phylogeny?** and (ii) **what predicts the evolution**
87 **of high fidelity in wasp mimics?** We utilise a ‘distance transform’ method for
88 image analysis (Taylor *et al.*, 2013) to quantify the similarity of Holarctic hoverflies
89 from 108 genera to the common and widespread social wasp model, *Vespula*
90 *germanica*. The distance transform approach allows rapid semi-automated

91 evaluation of mimetic accuracy across large numbers of taxa, which can easily
92 be re-run with different sub-sets of data, model taxa etc. We focus on wasp
93 mimicry because it is the most widespread form of mimicry in hoverflies, likely to
94 be homologous across species, and most easily quantified using our objective
95 image analysis. Having verified that our measure of similarity correlates well with
96 existing measures and similarity scores for two additional social wasp models,
97 we then plot pattern similarity onto the hoverfly phylogeny, and test for
98 associations with key life history traits. For the first time in a study of this kind, we
99 include hoverflies that are not considered to be mimics, so that we can identify
100 the positions in the phylogeny where wasp mimicry first evolved.

101 **2 Methods**

102 *2.1 Hymenopteran model selection*

103 We chose to study mimicry of the German wasp (*Vespula germanica*), a
104 widespread and abundant noxious social wasp considered to be a model for
105 many hoverfly mimics in the Holarctic region (Gilbert, 2005). *V. germanica* is very
106 similar in appearance to other *Vespula* species (Table S1; see Section 2.8), which
107 are also likely models for hoverfly mimicry, but *V. germanica* is the most widely
108 distributed and the most common species in the genus (CABI, 2019). Our specific
109 objective was to study the evolution of social wasp mimicry alone, rather than all
110 forms of Batesian mimicry in hoverflies. Where we find a hoverfly species is a
111 poor wasp mimic, or a non-mimic relative to wasps, this could be because it is
112 entirely non-mimetic, but it could also be because it is a conspicuous mimic of
113 another defended model. Other relevant putative models for hoverfly mimics
114 include honeybees (*Apis mellifera*) and bumblebees (*Bombus spp.*).

115 2.2 *Image selection*

116 We used images of hoverfly abdomens to characterise mimetic accuracy.
117 Logistical constraints, including a shortage of high-quality images, meant we
118 could not sample all Holarctic hoverfly species. As the species in most hoverfly
119 genera/subgenera have similar colour patterns, we chose a single representative
120 species from each for analysis (see supplementary data). If many species looked
121 similar to the human eye, the one with a distribution that most widely overlapped
122 with that of *V. germanica* was included. Where species had similar distributions,
123 the most abundant species (according to expert opinion, see below) was
124 included. Some genera/subgenera (25 out of 108) contained several widely-
125 distributed, abundant species with conspicuously different abdominal patterns. In
126 these cases, we included multiple representative species, one for each obvious
127 type of pattern, except where good quality images were unavailable. Thus, the
128 taxonomic units used here are colour-pattern groups usually corresponding to
129 genera or subgenera, but occasionally to species-groups within them (Table S2):
130 we use the term 'operational taxonomic unit' (OTU) to denote these groups. For
131 the full list of species used, see the supplementary dataset.

132 Hoverfly and wasp images were sourced primarily from reliable internet sites run
133 by taxonomic experts where species identification was judged to be accurate by
134 the research community (see supplementary data). Multiple images were
135 sourced from Taylor *et al.* (2017) and Speight and de Courcy Williams (2018).
136 Images were selected following a hierarchy of rules for quality, sexual dimorphism
137 and intraspecific variation. To meet the criteria for quality, the images were of
138 alive or recently dead specimens to avoid colour fading, except *Chrysosyrphus*

139 *nasuta* which, due to a lack of good images, was from an artist's drawing. The
140 images we used had variable backgrounds, depending on how the image was
141 acquired, so we ran Wilcoxon test comparing mimetic accuracy between images
142 from natural and artificial backgrounds to ensure our results were not impacted
143 by the image sources.

144 The abdomen was used for analysis because the colour pattern is, in general,
145 much more distinctive and variable on the abdomen than on the thorax in
146 dipterans and hymenopterans (Marchini *et al.*, 2017), and the abdomen is
147 typically conspicuous to potential predators. Studies have previously shown that
148 abdominal colour patterns of both hoverflies and wasps consist of clearly
149 delineated contrasts in both achromatic and chromatic dimensions, and do not
150 contain hidden ultra-violet signals (Taylor *et al.*, 2016b), meaning that the spatial
151 elements of the pattern are visible to all but the most primitive of visual systems.

152 Images were only used where they showed a clear dorsal view of the abdomen,
153 without obvious distortion of the pattern. Images with glare, reflections and
154 obstructions from pollen or wings were rejected unless no alternative was
155 available. Where the best image included minor examples of such imperfections,
156 these were corrected by eye in the image pre-processing stage using ImageJ
157 (Abràmoff *et al.*, 2004), for example by exploiting symmetry of the pattern to fill in
158 obscured areas. It is important to note that, since we relied on photographs in the
159 public domain, the selection of images we used was probably not entirely
160 representative of natural inter- and intraspecific variation. Photographs of larger,
161 more brightly-coloured species or individuals, and those with striking patterns,
162 are probably more likely to be made available in the sources we used, because

163 they are easier to photograph, or more interesting or detectable to photographers
164 and entomologists.

165 Images of males were used by default, except where images of females were of
166 significantly higher quality. Most of the selected species were not conspicuously
167 sexually dimorphic. There were four instances where females had to be chosen
168 despite the presence of conspicuous sexual dimorphism, defined as a distinct
169 difference in pattern markings not simply due to differences in the shape or size
170 of the abdomen: *Baccha elongata*, *Hiatomyia willistoni*, *Mixogaster breviventris*
171 and *Nausigaster punctulata*. Some multivoltine hoverflies, especially *Eristalis*
172 spp., exhibit phenotypic variation in colour pattern due to seasonal variation, so
173 an image of the most commonly recorded pattern was selected for analysis
174 (Holloway *et al.*, 1997). *Merodon equestris*, a bumblebee mimic, was not included
175 because it has widely variable and distinct colour morphs (Mengual *et al.*, 2006).

176 2.3 *Phylogeny reconstruction*

177 Recently, much progress has been made in our understanding of hoverfly
178 phylogeny at the genus level (Mengual *et al.*, 2018; Pauli *et al.*, 2018; Moran and
179 Skevington, 2019; Moran *et al.*, 2021), but its overall architecture remains little
180 changed from the study of Rotheray & Gilbert (1999) as modified by Ståhls *et al.*
181 (2003). We used a phylogeny based on morphological data from Katzourakis *et*
182 *al.* (2001), excluding non-Holarctic genera and a few that lack good quality
183 images. This phylogeny is in turn based on Rotheray and Gilbert's (1999, 2008)
184 cladistic study of larval characters in Palaearctic genera, and is very similar to
185 recent skeleton trees based on transcriptomics (Pauli *et al.*, 2018) and anchored
186 enrichment genetic data (Young *et al.*, 2016). A comprehensive phylogeny from

187 anchored enrichment data is currently being constructed, but is still a long way
188 from publication (JH Skevington, pers. comm.).

189 The Katzourakis *et al.* (2001) tree was updated using more recent molecular
190 phylogenies of restricted subgroupings and seventeen extra OTUs were added;
191 if no data on their placement were available, the relationship was left as a
192 polytomy (see Table S2). Our semi-resolved, literature-based tree was formed
193 using Mesquite (Version 3.6, Maddison and Maddison, 2018). In the absence of
194 a comprehensive resolved phylogeny, combining published trees is often better
195 than, for example, estimating the phylogeny using proxies from DNA sequences
196 in GenBank (Beaulieu *et al.*, 2012) and leaving parts unresolved where molecular
197 data are not available. Phylogenies which covered most of the species used in
198 this study took precedence over less densely sampled studies. Trees
199 extrapolated from model-based approaches, such as Bayesian and maximum
200 likelihood, took priority over those inferred from distance-based methods or
201 parsimony (Beaulieu *et al.*, 2012). These published data were used to resolve as
202 much of the tree as possible to create a 'master tree', which was then imported
203 into R version 3.5.2 (R CoreTeam, 2018) for analysis using the packages *ape*
204 (Paradis and Schliep, 2019) and *geiger* (Harmon *et al.*, 2007). Branch lengths
205 were calculated using the 'Grafen' algorithm, where the depth of nodes is equal
206 to the number of daughter species descend from that node (Grafen 1989), and
207 polytomies were made dichotomous (with zero length) using the 'compute.brlen'
208 and 'multi2di' functions in the *picante* package (Kembel *et al.*, 2010). The final
209 tree was constructed and visualised using *RColorBrewer* (Neuwirth and
210 Neuwirth, 2011) and the 'contMap' function in *phytools* (Revell, 2012).

211 2.4 *Image preparation*

212 Following image selection, three wasp species and a total of 167 OTUs within
213 108 genera of Holarctic hoverflies were selected for processing and analysis
214 (see supplementary dataset). Image pre-processing was carried out in ImageJ.
215 Firstly, images were rotated so that the top of the scutellum was horizontal, with
216 the tip of the abdomen facing downwards. Images were cropped to the smallest
217 area containing the abdomen, from the tip of the abdomen to where the
218 scutellum meets the two sides (Taylor *et al.*, 2013). Without changing the
219 aspect ratio, each image was scaled to the height of 100 pixels to standardise
220 abdomen size and the abdomen outlined in blue (Figure 1). In all cases, we
221 were able to identify two distinct colours in the abdominal pattern: a pale colour
222 (typically yellow, white or orange) and a dark background colour (typically black
223 or dark brown). Images were ‘segmented’ based on their light and dark
224 components using colour thresholding and paintbrush tools. Whilst in most
225 cases, the colour pattern was formed by pigmentation of the tergites, coloured
226 hairs sometimes played a role. The hairs outside the true outline of the
227 abdomen were only included if they were dense enough to 1) obscure the true
228 outline or 2) form a border just as strong as the true outline. Hairs within the
229 outline of the abdomen were only included if they would be conspicuous
230 regardless of the strength or direction of any light. All 167 images were pre-
231 processed, saved as TIFF files, and converted into a binary format using
232 MATLAB (Figure 1; Taylor *et al.*, 2013; MATLAB, 2018).

233 2.5 *Similarity calculation*

234 A matrix of dissimilarity values was produced in MATLAB according to the
235 methods in Taylor *et al.* (2013). To avoid misalignment and optimise the
236 dissimilarity value, the ‘optim’ parameter was set to ‘hy’ and the ‘scal’ parameter
237 was set to ‘y’. This shifted each image vertically to minimise mismatch between
238 segments whilst keeping the height and aspect ratio the same (Taylor *et al.*,
239 2013). To ease interpretation, results were scaled based on the highest number
240 in the matrix, converted to similarity values and squared. Henceforth, these
241 values are referred to as “distance transform similarity scores”. Images from non-
242 mimetic species with entirely black abdomens were assigned the similarity value
243 of zero. The ancestral estimates for similarity were calculated using the ‘fastAnc’
244 function from *phytools*, which assumes a Brownian model of evolution (Revell,
245 2012).

246 2.6 *Other measures of mimetic fidelity*

247 We used classifications of mimicry from several sources to calibrate the measure
248 of mimetic accuracy from our image analysis, and to establish a formal method
249 for categorising an OTU as a mimic. The calibration allowed us to determine
250 whether our similarity measure actually predicts the behaviour of representative
251 vertebrates (humans and birds) when faced with a visual discrimination task
252 similar to that required to identify models and mimics in real populations. First,
253 we collected expert evaluations of mimetic accuracy from the literature (Gilbert,
254 unpublished data collated over the past 40 years from ca. 10,000 syrphid
255 publications). Three categories were recognised: any OTU identified as a social
256 wasp mimic was labelled either ‘good’ or ‘poor’, based on the expert descriptions

257 given, whilst it was considered a ‘non-mimic’ when there was no source to say
258 otherwise.

259 Next, we gathered independent estimates of mimetic accuracy for a subset of
260 overlapping OTUs from published studies of pigeon (Dittrich *et al.*, 1993) and
261 human (Penney *et al.*, 2012) evaluations of hoverfly images. To increase
262 coverage to all 167 OTUs in our dataset, we also designed our own survey using
263 human volunteers. In contrast to the published studies mentioned above, which
264 evaluated full-colour images of the whole hoverfly, we surveyed perceptions of
265 wasp mimicry in the binary images of abdomens created for the distance
266 transform analysis. This permitted direct comparison of human perception of
267 mimetic accuracy and distance transform similarity scores, based on the same
268 characters. Non-expert volunteers were recruited from a student population and
269 were asked to compare the abdomen patterns of *V. germanica* and each of 30
270 hoverfly OTUs, randomly selected without replacement from the pool of 167
271 images. Volunteers rated the similarity of the pair of images from 1 (hoverfly is
272 not mimetic) to 10 (perfect mimicry). Each pair of images was displayed via a
273 website on the volunteer’s computer screen until they decided on a rating and
274 clicked the button. Overall, the survey was completed 98 times, and each image
275 was assessed a minimum of 8 and a maximum of 29 times.

276 2.7 Analyses

277 All statistical analyses were conducted in R, version 3.5.2 (R CoreTeam, 2018).
278 Phylogenetic Generalised Least Squares (PGLS) analyses were performed using
279 the *caper* package to investigate the relationship between pattern similarity and
280 ecological characteristics whilst correcting for phylogenetic effects (Orme *et al.*,

281 2018). These traits included larval feeding ecology, voltinism, phenology (mostly
282 from Speight, 2018) and, as a proxy for body size, wing length (taken from Gilbert,
283 unpublished data (see above); Stubbs and Falk, 2002). The key flight periods
284 were defined as 'early' (March to May), 'mid' (May to July) and 'late' (July to
285 September), based on quantitative data (primarily the Hoverfly Recording
286 Scheme www.hoverfly.org.uk, with gaps filled from Gilbert, unpublished data, see
287 above). The PGLS approach considered the absence of phylogenetic
288 independence between these traits by incorporating a covariance matrix between
289 species into the model. Phylogenetic signal in the model was measured using a
290 maximum likelihood estimation of the parameter lambda (Pagel, 1999), which
291 varies from zero (phylogenetic independence of residuals) to one (strong
292 association of residuals with phylogeny under the Brownian motion model of
293 evolution). We estimated the degree of phylogenetic signal in the individual traits
294 measuring mimicry (both the distance transform scores, and the human
295 evaluation scores), by fitting intercept-only models predicting both traits.

296 PGLS analyses were performed using all ecological traits as explanatory
297 variables, using similarity scores from the distance transform analysis (one for
298 each wasp model) and our survey as separate response variables. Typically, it is
299 not necessary to carry out non-phylogenetically-controlled analyses in addition to
300 PGLS (Freckleton, 2009), but since there is some uncertainty over the phylogeny
301 used, we also modelled the data using ordinary least-squares (OLS) regression.
302 Models with the best fit were identified using stepwise model selection based on
303 Akaike's information criterion (AIC). This involved starting with the full model

304 containing the complete set of predictors, then sequentially removing the least
305 significant variable one at a time to find which model had the lowest AIC value.

306 To explore the impact of considering social wasp mimicry as a discrete as
307 opposed to a continuous trait, we inspected the distribution of our *V. germanica*
308 distance transform similarity scores for each category from the literature and
309 identified a threshold score below which there are no recognised mimics (Figure
310 2). We used this threshold to create a variable for mimicry as a binary trait (1/0).
311 As a large number of hoverflies above this threshold were classified as non-
312 mimics in the literature, we ran a second binary analysis where the threshold was
313 defined by the point above which the number hoverflies classified as mimics by
314 the literature exceeded the number of non-mimics. We also evaluated binary
315 mimicry using the raw data from the literature evaluation, where 'good' and 'poor'
316 mimics were grouped together under 'mimics' and compared with OTUs for which
317 no mimicry was reported. These three definitions of binary mimicry are
318 subsequently referred to as 'the mimicry threshold', 'the majority threshold' and
319 'the literature categories' respectively. For each definition of binary mimicry, a
320 phylogenetic logistic regression was performed using the 'phyloglm' function in
321 *phyloglm*, which uses alpha (α) to represent the strength of the phylogenetic
322 signal (Ives and Garland, 2009). A low alpha value denotes a strong association
323 between phylogenetic structure and trait presence. Models in the phyloglm
324 analysis were compared using AIC.

325 2.8 Sensitivity tests

326 We ran a supplementary analysis using two additional social wasp models,
327 *Vespula vulgaris* (the second most common member of the genus) and *Polistes*

328 *dominula* (another widespread and common social wasp), to establish how
329 sensitive our findings were to the choice of model taxon.

330 Our approach to image analysis is less effective where aposematic and mimetic
331 patterns on the abdomen rely on coloured hairs, as is the case with bees and
332 some of their mimics, because the abdominal patterns of hairy species do not
333 have uniform patches of colour. In the distance transform algorithm, this leads
334 to abnormally high similarity values when compared to a wide range of possible
335 patterns, since the distances between matching pixels are small. Hence, we
336 were unable to extend our analysis to include bee mimicry. For some hairy
337 species the distance transform measure of mimetic accuracy did not correspond
338 well with evaluations of wasp mimicry made by volunteers or the literature (see
339 section 3.3). We therefore explored the impact of the inclusion of hairy species
340 in the dataset by classifying each species as hairy (with conspicuous hairs on
341 the abdomen, $n = 32$) or not hairy ($n = 135$), and including this as a factor in the
342 analysis of the relationship between the distance transform score and similarity
343 to *V. germanica* as perceived by our volunteers. We also ran a supplementary
344 phylogenetic analysis for *V. germanica* distance transform similarity scores
345 without the hairy species included.

346 We were concerned about the influence of sampling bias in the estimation of
347 phylogenetic signal in our main analysis, caused by the repeated sampling of
348 some genera in which phenotypes varied conspicuously among species (see
349 above). We therefore conducted a second analysis with a reduced version of our
350 *V. germanica* dataset. We repeated the PGLS and *phyloglm* binary analysis 1000
351 times with just one randomly selected species from each genus in which we had

352 data for multiple species, and generated Higher Posterior Density (HPD)
353 confidence intervals for the model coefficients averaged across all 1000 trees.

354 **3 Results**

355 3.1 Quantifying mimetic similarity

356 Abdominal pattern similarity of hoverflies to *V. germanica* was widely distributed
357 (Figure S1). The distance transform analysis identified the three best *V.*
358 *germanica* mimics as *Spilomyia interrupta*, *Caliprobola speciosa* and *Helophilus*
359 *pendulus* (Figure S2). Aside from the all-black species, the three lowest similarity
360 scores were obtained from *Hadromyia grandis*, *Pyrophaena rosarum* and
361 *Volucella pellucens* (Figure S2). This result was the same in our analysis
362 excluding species with hairy abdomens, but the choice of model taxon had some
363 impact on the ranking of the mimics (Table 1; Figure S2). Nevertheless, the
364 similarity scores in relation to *V. vulgaris* (Spearman's rank: $r_s = 0.83$, $p < 0.001$)
365 and *P. dominula* (Spearman's rank: $r_s = 0.78$, $p = < 0.001$) were strongly and
366 significantly correlated with those for *V. germanica* (Figure S3). The similarity
367 scores of every hoverfly species in relation to all three wasp models are provided
368 in the supplementary dataset. The image background, and therefore the image
369 source, did not impact the similarity score (Wilcoxon rank sum test: $W = 2729.5$,
370 $p = 0.17$). Inspection of the distribution of distance transform similarity scores for
371 species classified as mimics in the literature suggested a threshold of 0.74, below
372 which hoverflies are never considered to be social wasp mimics (Figure 2). This
373 threshold was the same when species with hairy abdomens were excluded
374 (Figure S4). The threshold above which the majority of species were considered
375 mimics by the literature was 0.808 (number of mimics above threshold = 45;

376 number of non-mimics above threshold = 42; Figure 2). These two thresholds
377 were used to divide mimics from non-mimics for subsequent analyses.

378 3.2 *Distance transform and previous studies*

379 Our distance transform similarity scores for hoverflies differed significantly across
380 descriptive categories found in the literature (see section 2.6 above), with ‘non-
381 mimics’ having the lowest similarity to *V. germanica* (Kruskal-Wallis: Chi-squared
382 = 52.83, $df = 2$, $p < 0.001$). Although the difference between ‘good’ and ‘poor’
383 mimics was not significant (Dunn’s test: $z = 1.07$, $p = 0.14$), ‘good’ mimics were
384 marginally more similar on average (Figure S5). The results when hairy species
385 were excluded were qualitatively similar (Figure S6). Distance transform similarity
386 scores were significantly positively correlated with similarity analyses from
387 published studies of pigeon (Spearman’s rank: $r_s = 0.73$, $p = 0.02$; Figure S7A;
388 Dittrich *et al.*, 1993) and human perception ($r_s = 0.74$, $p = 0.0002$; Figure S7B;
389 Penney *et al.*, 2012).

390 3.3 *Our survey*

391 Volunteer perception of wasp mimicry in binary images of hoverfly abdomens in
392 our survey varied significantly between mimics and ‘non-mimics’, as defined by
393 the literature (Kruskal-Wallis: Chi-squared = 57.89, $df = 2$, $p < 0.001$), but not
394 between ‘good’ and ‘poor’ mimics (Dunn’s test: $z = 0.84$, $p = 0.20$; Figure S8).
395 The average perceived similarity in our survey was also positively correlated with
396 survey ratings from Penney *et al.* (2012) (Pearson’s correlation coefficient: $r =$
397 0.86 , $p < 0.001$; Figure S9). The ranking of distance transform similarity scores
398 was also negatively correlated with the survey results ($r_s = -0.75$, $p < 0.001$;

399 Figure S10) – species with a higher similarity score in the distance transform
400 analysis were typically perceived to be more similar to *V. germanica* in our survey.
401 Many of the species with hairy abdomens appeared to be outliers, with a low
402 survey score but relatively high distance transform similarity ranking (Figure S10).
403 A two-way ANOVA indicated that hairy species have significantly higher distance
404 transform similarity ranks overall ($F_{(1,164)} = 281.63$; $p < 0.001$), and their
405 relationship with survey score is weaker, though not significantly so ($F_{(1,163)} =$
406 3.266 ; $p = 0.073$). The results of subsequent sensitivity tests where species with
407 hairy abdomens were excluded from the *V. germanica* dataset are summarised
408 in Table 1.

409 3.4 *The evolution of mimicry*

410 Social wasp mimicry, as revealed by distance transform analysis of hoverfly
411 abdominal patterns, was patchily distributed over the phylogeny (Figure 3). When
412 we defined species as mimics or non-mimics by calibrating similarity scores using
413 the literature (see above), transitions between states of mimicry appear to have
414 happened repeatedly, both from non-mimetic to mimetic and vice versa.
415 According to ancestral state estimations using our mimicry threshold of 0.74,
416 *Vespula germanica* mimicry has evolved 35 times, 13 of these being at ancestral
417 nodes (47 and 16 times respectively using the majority threshold of 0.808) and
418 there were seven instances (twelve using the majority threshold, three of these
419 being at ancestral nodes; Figure S11) where non-mimics were found deep within
420 a clade of mimics (Figure 3). When binary mimicry was defined by the literature
421 evaluation, mimicry evolved 28 times, nine of these being at shared ancestral
422 nodes (Figure 3). The Pipizinae were all non-mimics, whereas Eristalinae and

423 Syrphinae contained species which were quite variable in their mimetic accuracy.
424 Microdontinae, the earliest evolving subfamily, had high similarity results and
425 therefore the two species we examined were considered to be accurate mimics
426 of *V. germanica*. The pattern of repeated evolution of mimicry was broadly similar,
427 regardless of the choice of wasp model (Figure S13).

428 The phylogenetic signal associated with the distance transform similarities to *V.*
429 *germanica* was significantly different from zero, but not strong, because the
430 observed value was also significantly different from one ($\lambda = 0.63$, 95% CI = 0.59
431 $- 0.81$, $p(\lambda = 0) < 0.001$, $p(\lambda = 1) < 0.001$). The same was true for both *V. vulgaris*
432 and *P. dominula* (Figure S13), but the phylogenetic signal was slightly weaker in
433 analyses of the *V. germanica* similarity survey ($\lambda = 0.43$, 95% CI = 0.15 $-$ 0.71, p
434 ($\lambda = 0) < 0.001$, $p(\lambda = 1) < 0.001$) and sensitivity tests (Table 1).

435 3.5 *Life history correlates of mimicry*

436 The fit of the PGLS models was better than equivalent OLS models for all three
437 wasp species, which establishes that the evolution of mimicry is constrained by
438 phylogeny (see Table S3). The best statistical models for the distance-transform
439 scores for each wasp and the survey similarity scores for *V. germanica* all
440 revealed that the most significant variables explaining mimetic similarity were
441 wing length and voltinism (Table 1; Table 2). Smaller species were significantly
442 less mimetic than larger species (Figure S14) and univoltine species were
443 significantly worse mimics than multivoltine species, with bivoltine somewhere
444 between the two (Figure S15). There were no noticeable relationships between
445 mimicry and larval feeding ecology (Figure S16). Species which emerge later in
446 the year were typically slightly better mimics, but this effect of phenology was not

447 significant (Figure S16). Our analysis of *V. germanica* mimicry as a binary trait
448 showed qualitatively similar results, with the results varying to some extent
449 depending on which species were selected when reanalysing the data excluding
450 all but one species per genus (Table S6; Table S7).

451

452 **4 Discussion**

453 Our study provides the first systematic and quantitative description of the
454 repeated evolution of social wasp mimicry across the entire Holarctic hoverfly
455 family. Distance transform analysis of abdominal patterns provides a measure of
456 mimetic accuracy which can be applied to large numbers of taxa simultaneously
457 and is not tied to a particular visual system. Our results show that this measure
458 strongly corroborates other assessments of mimetic accuracy from expert and
459 non-expert humans and birds, and extends our understanding of variation in
460 abdominal patterns to species for which wasp mimicry has not previously been
461 evaluated, or has been considered to be absent. We found that accurate wasp
462 mimicry has probably evolved repeatedly in hoverflies, and may also have been
463 lost. We also found that mimetic accuracy is predicted by life history: it correlates
464 positively with a proxy for body size, and is associated with voltinism. This implies
465 that hoverfly ecology influences the tendency for species to evolve wasp mimicry
466 (or indeed the reverse), giving us an insight into origins of the tremendous
467 variation in morphology we see across the family.

468 Our results suggest social wasp mimicry has evolved repeatedly at scattered
469 positions throughout the phylogeny, regardless of which threshold we use to

470 distinguish between mimics and non-mimics. The phylogenetic signal for wasp
471 similarity was significant but not strong, suggesting some relationship between
472 evolutionary history and mimetic fidelity, but with some lability. Similarity to *V.*
473 *germanica* in the most basal of the taxa used, *Mixogaster* and *Microdon*, indicates
474 that mimicry evolved early. However, this is a very provisional result because we
475 could only sample two species of this very diverse predominantly Neotropical
476 subfamily (552 species, Reemer and Stahls 2013a). Despite this, the deepest
477 nodes had similarity estimates lower than our threshold, suggesting that the basal
478 character state for the Syrphidae was non-mimicry of wasps, and that our
479 Microdontinae may not appropriately represent the ancestral phenotype
480 (although one of them, *Mixogaster*, is thought to be basal amongst the
481 Microdontinae: Reemer and Stahls, 2013b).

482 Our results suggest that wasp mimicry has occasionally been lost deep within a
483 clade of good wasp mimics; thus, to assume that conspicuous wasp-mimetic
484 hoverflies always evolve from non-mimetic ancestral phenotypes may be
485 inappropriate (Figure 3; see also Kikuchi and Pfennig, 2010; Hodson and
486 Lehtinen, 2017). The loss of mimetic accuracy could result from an alteration in
487 the selective environment which meant that wasp mimicry was no longer an
488 advantageous adaptation. For example, none of the ecological traits examined
489 for *Leucozona lucorum* were noticeably different relative to its closely related
490 taxa, so one possible explanation for the loss of mimetic resemblance to wasps
491 could be a change in hymenopteran model. *L. lucorum* has been described as ‘a
492 little bumblebee-like’, unlike closely related taxa which have been identified more
493 with mimics of social and solitary wasps (Röder 1990). This supports the

494 conclusion that additional research on the similarity between hoverflies and other
495 models is needed to understand the evolution of this multifaceted trait fully (see
496 below).

497 In all our analyses, wing length was a good predictor of wasp mimicry (Table 1;
498 Table 2). Larger species were typically better wasp mimics, which corresponds
499 with experimental results and theoretical hypotheses from previous papers
500 (Penney *et al.*, 2012; Taylor *et al.*, 2016a). There may be greater selection
501 pressure on larger hoverflies to deceive predator visual systems because they
502 are more nutritionally profitable prey items (Penney *et al.*, 2012). Smaller
503 hoverflies also take longer to warm up to flight temperatures (Morgan and
504 Heinrich, 1987), potentially increasing the thermoregulatory cost of pale colours
505 (Taylor *et al.*, 2016a), since darker colours allow hoverflies to warm up more
506 rapidly (Holloway *et al.*, 1997). Thus, thermoregulatory costs might act in
507 opposition to selection for accurate mimicry, especially in smaller species. This
508 is demonstrated by the 26 species with entirely black abdomens, which all had
509 wing lengths below 10mm (Figure S14). Alternatively, small size may enable
510 predators to discriminate prey from models, and hence there is no benefit for a
511 small species evolving to be mimetic.

512 Voltinism was also an explanatory variable for pattern similarity. Multivoltine
513 species had significantly more similar abdomen patterns to *V. germanica*, and
514 were therefore better mimics, than univoltine species, with bivoltine species being
515 intermediate (Table 2). More generations per year may lead to better mimicry
516 because there are more chances for selection to act in a given time frame
517 (Gillman and Wright, 2014). Furthermore, univoltine species emerge at a

518 particular time of year for a relatively short time, and if this does not coincide with
519 a high abundance of models there may be less selection for good mimicry
520 (Howarth and Edmunds, 2000; Finkbeiner *et al.*, 2018; Hassal *et al.*, 2019).
521 Multivoltine species are essentially present all year round, and so are bound to
522 coincide with the peaks of wasp abundance in spring, when queens search for
523 nests, and late summer when the nest is at maximum size (Tryjanowski *et al.*,
524 2010). Although phenology was not a significant predictor of wasp mimicry (Table
525 S3), results suggest that the earliest emerging species could in general be the
526 weakest mimics, which is somewhat consistent with this hypothesis (Figure
527 S16B).

528 The selection and definition of traits for study by evolutionary biologists is always
529 influenced by human perception, and is by necessity somewhat arbitrary. To the
530 human eye, mimicry is clearly present in some hoverflies, and absent in others,
531 but studying this variation scientifically requires us to define the trait more
532 precisely, answering questions about sensory modality (e.g., are we considering
533 only visual mimicry?), specificity (e.g., are we considering mimicry of one model
534 species or several?), and variability (e.g., is mimicry a quantitative or discrete
535 trait)? By choosing to study similarity to the abdomen pattern of *Vespula*
536 *germanica*, we were able to make considerable progress in quantifying variation
537 in mimicry across the hoverflies. Interestingly, despite the variable approach to
538 the characterisation of mimicry in the literature, our tightly-defined quantitative
539 measure typically corresponded very well with more subjective evaluations from
540 other published studies. The correspondence was not perfect, however, and the
541 descriptive classification of hoverflies as “good” or “poor” mimics in particular was

542 not a strong predictor of our similarity scores. The failure to differentiate between
543 good and poor mimics may either be because humans perceive mimicry in a fairly
544 binary manner, or because the classification into “good” and “poor” in the
545 literature has not been made in a consistent or systematic way.

546 By comparing two different benchmarks for wasp mimicry to how it is categorised
547 by the literature, we aimed to gain insight into the effects of different methods for
548 defining mimicry as a discrete trait. Figure 2 and our binary analyses highlight
549 how wasp mimicry is more of a continuous spectrum than a binary, or categorical
550 trait, which has important implications for how future studies define mimicry. It is
551 also important to note the majority threshold for mimicry was still passed by 52%
552 of the hoverflies studied here, suggesting that wasp mimicry could be a much
553 more prevalent feature of natural communities than previously estimated (22%:
554 Gilbert, 2005; Kikuchi *et al.*, 2021). Even the vaguest resemblance to a noxious
555 or abundant model can afford protection to a mimic, perhaps because the optimal
556 predator behaviour may be to avoid risks by not sampling even poor mimics
557 whenever possible, resulting in relaxed selection on mimetic accuracy (Gilbert,
558 2005; Pfennig and Kikuchi, 2012; Sherratt and Peet-Paré, 2017). Just as
559 Nicholson (1927) claimed almost 100 years ago, our results suggest that the
560 literature may have underestimated the amount of mimicry in nature, potentially
561 by overestimating the gap in predation pressure among mimics (Dittrich *et al.*,
562 1993).

563 An alternative explanation for our apparent detection of previously undescribed
564 mimics is that the taxa with intermediate accuracy (in Figure 2) may actually have
565 abdomens which are never perceived to be mimetic by predators. The subjective

566 evaluations of wasp mimicry from the literature were typically made on the basis
567 of the entire appearance, and possibly even the behaviour, of the organism.
568 Some species with non-mimetic abdomens may thus be regarded as mimics for
569 other reasons, and this may mean that the thresholds we used (in Figure 2) are
570 poorly positioned to define abdominal pattern mimicry. Additionally, a taxon was
571 defined as a 'non-mimic' of *V. germanica* when there was no literature to say
572 otherwise, but many of these taxa were reported to be good mimics of other
573 models which themselves resemble wasps. For example, the 'non-mimic'
574 *Microdon analis* has been described as a good honeybee mimic (Röder, 1990)
575 but also received a high similarity score when compared to *V. germanica*.
576 Essentially, the overshadowing by more obvious putative models has contributed
577 to the inconclusive definition of Batesian mimicry (Gilbert, 2005). Evidently,
578 subjective literature assessments are not a reliable source for defining mimetic
579 accuracy.

580 The evaluation of mimicry as a trait is complicated considerably by the choice of
581 model taxon with which putative mimics are compared. If similarity scores were
582 high for several different models, this could be evidence for the multi-model
583 hypothesis, whereby some mimetic phenotypes are predicted to be an optimal
584 intermediate between several aposematic models (Edmunds, 2000; Sherratt,
585 2002). However, mimicry of animals as different as bumblebees and social wasps
586 can involve very different morphological (and other – e.g., behavioural, or
587 perhaps even acoustic) characters, presumably encoded by different sets of
588 genes. If we want to explore the pattern of selection on mimicry across the
589 phylogeny, it seems sensible to start by focusing on a more narrowly defined trait,

590 where it is likely the mimetic phenotypes exhibited by different species are mostly
591 homologous. So, we chose to examine visual mimicry of the social wasp *V.*
592 *germanica*. *V. germanica* is the most common and widespread species of social
593 wasp across the Holarctic, so it provides a reasonable best guess at the
594 phenotypic target for selection on this form of mimicry. Our results were largely
595 insensitive to this choice: hoverfly similarity to two other social wasps (*V. vulgaris*
596 and *P. dominula*) showed similar patterns across the phylogeny, and similar
597 associations with life history traits. A fascinating unanswered question is how
598 social wasp mimicry in hoverflies is related to mimicry of other Hymenoptera. For
599 example, to what extent were the genes and corresponding phenotypes involved
600 in wasp mimicry co-opted in honeybee or even bumblebee mimicry (or vice versa)
601 during diversification of the lineage? Are the different forms of mimicry seen in
602 hoverflies, corresponding to different model taxa, driven by similar predators, and
603 associated with similar life history traits? Only by addressing these questions with
604 further research will we understand the extent to which it is reasonable to
605 consider hoverfly mimicry of any hymenopteran to be a meaningful single trait.

606 This research has provided insights into the ecological and evolutionary factors
607 that shape complex phenotypes by advancing our understanding of mimetic
608 pattern evolution in a well-studied Batesian system (Penney *et al.*, 2012; Kikuchi
609 and Pfennig, 2013; Marchini *et al.*, 2017). Our results suggest that wasp mimicry
610 is a relatively labile trait which has evolved repeatedly, and that this is at least
611 partly predictable from life history. Since these conclusions apply specifically to
612 the hoverfly abdomen in its visual mimicry of social wasps, further work is needed
613 to explore the extent to which different forms of mimicry (e.g., toward other model

Mapping the evolution of accurate mimicry

614 Hymenoptera, and in other sensory modalities) show similar patterns of evolution.

615 It is clear to us, however, that objective phylogenetically-controlled comparative

616 studies of mimicry continue to illuminate the selective forces which shape the

617 evolution of phenotypes in natural populations.

618

619 **References**

- 620 Abràmoff, M.D., Magalhães, P.J. and Ram, S.J., 2004. Image processing with
621 ImageJ. *Biophotonics International*, 11(7), pp.36-42.
- 622 Azmeh, S., Owen, J., Sørensen, K., Grewcock, D. and Gilbert, F., 1998. Mimicry
623 profiles are affected by human-induced habitat changes. *Proceedings of*
624 *the Royal Society of London B: Biological Sciences*, 265(1412), pp.2285-
625 2290.
- 626 Bates, H.W., 1862. XXXII. Contributions to an insect fauna of the Amazon valley.
627 Lepidoptera: Heliconidæ. *Transactions of the Linnean Society of*
628 *London*, 23(3), pp.495-566.
- 629 Beaulieu, J.M., Ree, R.H., Cavender-Bares, J., Weiblen, G.D. and Donoghue,
630 M.J., 2012. Synthesizing phylogenetic knowledge for ecological
631 research. *Ecology*, 93(sp8), pp.S4-S13.
- 632 CABI, 2019. *Vespula germanica* distribution map. *Invasive Species*
633 *Compendium*. Wallingford, UK: CAB International. www.cabi.org/isc
- 634 Chandler, P.J., 1998. Checklists of insects of the British Isles. *Handbooks for the*
635 *Identification of British Insects*, 12, pp.1-234.
- 636 Cuthill, I.C. and Bennett, A.T., 1993. Mimicry and the eye of the
637 beholder. *Proceedings of the Royal Society of London. Series B:*
638 *Biological Sciences*, 253(1337), pp.203-204.
- 639 Dittrich, W., Gilbert, F., Green, P., McGregor, P. and Grewcock, D., 1993.
640 Imperfect mimicry: a pigeon's perspective. *Proceedings of the Royal*
641 *Society of London B: Biological Sciences*, 251(1332), pp.195-200.

- 642 Edmunds, M., 2000. Why are there good and poor mimics? *Biological Journal of*
643 *the Linnean Society*, 70(3), pp.459-466.
- 644 Edmunds, M. and Reader, T., 2014. Evidence for Batesian mimicry in a
645 polymorphic hoverfly. *Evolution*, 68(3), pp.827-839.
- 646 Finkbeiner, S.D., Salazar, P.A., Nogales, S., Rush, C.E., Briscoe, A.D., Hill *et al.*,
647 2018. Frequency dependence shapes the adaptive landscape of imperfect
648 Batesian mimicry. *Proceedings of the Royal Society of London B:*
649 *Biological Sciences*, 285(1876), p.20172786.
- 650 Freckleton, R.P., 2009. The seven deadly sins of comparative analysis. *Journal*
651 *of Evolutionary Biology*, 22(7), pp.1367-1375.
- 652 Gilbert, F., 2005. The evolution of imperfect mimicry. In: Fellowes, M., Holloway
653 and G., Rolff, J. (editors), *Insect Evolutionary Ecology*. CABI Publishing:
654 Wallingford, pp. 231-288.
- 655 Gillman, L.N. and Wright, S.D., 2014. Species richness and evolutionary speed:
656 the influence of temperature, water and area. *Journal of Biogeography*,
657 41(1), pp. 39-51.
- 658 Golding, Y.C., Edmunds, M. and Ennos, A.R., 2005. Flight behaviour during
659 foraging of the social wasp *Vespula vulgaris* (Hymenoptera: Vespidae)
660 and four mimetic hoverflies (Diptera: Syrphidae) *Sericomyia silentis*,
661 *Myathropa florea*, *Helophilus* sp. and *Syrphus* sp. *Journal of Experimental*
662 *Biology*, 208(23), pp.4523-4527.
- 663 Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of*
664 *the Royal Society of London. Series B, Biological Sciences*, 326, pp. 119-

665 157.

666 Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. and Challenger, W., 2007.
667 GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24(1),
668 pp.129-131.

669 Hassall, C., Billington, J. and Sherratt, T.N., 2019. Climate-induced
670 phenological shifts in a Batesian mimicry complex. *Proceedings of the*
671 *National Academy of Sciences*, 116(3), pp.929-933.

672 Hodson, E.E. and Lehtinen, R.M., 2017. Diverse evidence for the decline of an
673 adaptation in a coral snake mimic. *Evolutionary Biology*, 44(3), pp.401-
674 410.

675 Holen, Ø.H. and Johnstone, R.A., 2004. The evolution of mimicry under
676 constraints. *The American Naturalist*, 164(5), pp.598-613.

677 Holloway, G.J., Marriott, C.G. and Crocker, H.J., 1997. Phenotypic plasticity in
678 hoverflies: the relationship between colour pattern and season in
679 *Episyrphus balteatus* and other Syrphidae. *Ecological Entomology*, 22(4),
680 pp.425-432.

681 Holloway, G., Gilbert, F. and Brandt, A., 2002. The relationship between mimetic
682 imperfection and phenotypic variation in insect colour patterns.
683 *Proceedings of the Royal Society of London B: Biological Sciences*,
684 269(1489), pp.411-416.

685 Howarth, B. and Edmunds, M., 2000. The phenology of Syrphidae (Diptera): are
686 they Batesian mimics of Hymenoptera? *Biological Journal of the Linnean*
687 *Society*, 71(3), pp.437-457.

- 688 Howarth, B., Edmunds, M. and Gilbert, F., 2004. Does the abundance of hoverfly
689 (Syrphidae) mimics depend on the numbers of their hymenopteran
690 models? *Evolution*, 58(2), pp.367-375.
- 691 Ives, A.R. and Garland Jr, T., 2009. Phylogenetic logistic regression for binary
692 dependent variables. *Systematic Biology*, 59(1), pp.9-26.
- 693 Katzourakis, A., Purvis, A., Azmeh, S., Rotheray, G. and Gilbert, F., 2001.
694 Macroevolution of hoverflies (Diptera: Syrphidae): the effect of using
695 higher-level taxa in studies of biodiversity, and correlates of species
696 richness. *Journal of Evolutionary Biology*, 14(2), pp.219-227.
- 697 Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly,
698 D.D., Blomberg, S.P. and Webb, C.O., 2010. Picante: R tools for
699 integrating phylogenies and ecology. *Bioinformatics*, 26(11), pp.1463-
700 1464.
- 701 Kikuchi, D.W. and Pfennig, D.W., 2010. High-model abundance may permit the
702 gradual evolution of Batesian mimicry: an experimental test. *Proceedings
703 of the Royal Society of London B: Biological Sciences*, 277(1684),
704 pp.1041-1048.
- 705 Kikuchi, D.W. and Pfennig, D.W., 2013. Imperfect mimicry and the limits of natural
706 selection. *The Quarterly Review of Biology*, 88(4), pp.297-315.
- 707 Kikuchi, D.W., Herberstein, M.E., Barfield, M., Holt, R.D. and Mappes, J., 2021.
708 Why aren't warning signals everywhere? On the prevalence of
709 aposematism and mimicry in communities. *Biological Reviews of the
710 Cambridge Philosophical Society*.

- 711 Maddison, W. P. and D.R. Maddison, 2018. Mesquite: a modular system for
712 evolutionary analysis. Version 3.6. Available from:
713 <http://www.mesquiteproject.org>
- 714 Marchini, M., Sommaggio, D. and Minelli, A., 2017. Playing with black and yellow:
715 the evolvability of a Batesian mimicry. *Evolutionary Biology*, 44(1), pp.100-
716 112.
- 717 MATLAB, 2018. *MATLAB*. The Mathworks, Natick, Massachusetts, USA.
- 718 McLean, D.J., Cassis, G., Kikuchi, D.W., Giribet, G. and Herberstein, M.E., 2019.
719 Insincere flattery? Understanding the evolution of imperfect deceptive
720 mimicry. *The Quarterly Review of Biology*, 94(4), pp.395-415.
- 721 Mengual, X., Ståhls, G., Vujić, A. and Marcos-Garcia, M.A., 2006. Integrative
722 taxonomy of Iberian *Merodon* species (Diptera,
723 Syrphidae). *Zootaxa*, 1377, pp.1-26.
- 724 Mengual, X., Ståhls, G., Láska, P., Mazánek, L. and Rojo, S., 2018. Molecular
725 phylogenetics of the predatory lineage of flower flies *Eupeodes-Scaeva*
726 (Diptera: Syrphidae), with the description of the Neotropical genus
727 *Austroscaeva* gen. nov. *Journal of Zoological Systematics and*
728 *Evolutionary Research*, 56(2), pp.148-169.
- 729 Moore, C.D. and Hassall, C., 2016. A bee or not a bee: an experimental test of
730 acoustic mimicry by hoverflies. *Behavioral Ecology*, 27(6), pp.1867-1774.
- 731 Moran, K.M., Skevington, J.H., Kelso, S., Mengual, X., Jordaens, K., Young,
732 A.D., Ståhls, G., Mutin, V., Bot, S., van Zuijen, M. and Ichige, K., 2021. A
733 multigene phylogeny of the eristaline flower flies (Diptera: Syrphidae),

- 734 with emphasis on the subtribe Criorhinina. *Zoological Journal of the*
735 *Linnean Society*.
- 736 Moran, K.M. and Skevington, J.H., 2019. Revision of world *Sphecomyia* Latreille
737 (Diptera, Syrphidae). *ZooKeys*, 836, pp.15-79.
- 738 Nicholson, A.J., 1927. A new theory of mimicry in insects. *Australian Zoologist*,
739 5, pp.10-104.
- 740 Neuwirth, E. and Neuwirth, M.E., 2011. *Package 'RColorBrewer'*. CRAN 2011-
741 06-17 08: 34: 00. Apache License 2.0.
- 742 Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N. and
743 Pearse, W., 2018. caper: comparative analysis of phylogenetics and
744 evolution in R. *R Package Version 1.0.1*. [https://CRAN.R-](https://CRAN.R-project.org/package=caper)
745 [project.org/package=caper](https://CRAN.R-project.org/package=caper)
- 746 Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*,
747 401, pp.877–884.
- 748 Paradis E. and Schliep K., 2019. Ape 5.0: an environment for modern
749 phylogenetics and evolutionary analysis in R. *Bioinformatics*, 35, pp.526-
750 528.
- 751 Pauli, T., Burt, T.O., Meusemann, K., Bayless, K., Donath, A., Podsiadlowski, L.,
752 Mayer, C., Kozlov, A., Vasilikopoulos, A., Liu, S. and Zhou, X.I.N., 2018.
753 New data, same story: phylogenomics does not support Syrphoidea
754 (Diptera: Syrphidae, Pipunculidae). *Systematic Entomology*, 43(3),
755 pp.447-459.

- 756 Penney, H.D., Hassall, C., Skevington, J.H., Abbott, K.R. and Sherratt, T.N.,
757 2012. A comparative analysis of the evolution of imperfect mimicry.
758 *Nature*, 483(7390), pp.461 - 464.
- 759 Penney, H.D., Hassall, C., Skevington, J.H., Lamborn, B. and Sherratt, T.N.,
760 2014. The relationship between morphological and behavioral mimicry in
761 hover flies (Diptera: Syrphidae). *The American Naturalist*, 183(2), pp.281-
762 289.
- 763 Pfennig, D.W. and Kikuchi, D.W., 2012. Competition and the evolution of
764 imperfect mimicry. *Current Zoology*, 58(4), pp.608-619.
- 765 R Core Team, 2018. *R: A language and environment for statistical computing*. R
766 Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-](https://www.R-project.org/)
767 [project.org/](https://www.R-project.org/)
- 768 Reemer, M. and Ståhls, G., 2013a. Generic revision and species classification of
769 the Microdontinae (Diptera, Syrphidae). *ZooKeys*, 288, pp.1-123.
- 770 Reemer, M. and Ståhls, G., 2013b. Phylogenetic relationships of Microdontinae
771 (Diptera: Syrphidae) based on molecular and morphological
772 characters. *Systematic Entomology*, 38(4), pp.661-688.
- 773 Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology
774 (and other things). *Methods in Ecology and Evolution*, 3(2), pp.217-223.
- 775 Röder, G., 1990. *Biologie der Schwebfliegen Deutschlands (Diptera: Syrphidae)*.
776 E. Bauer.

- 777 Rotheray, G. and Gilbert, F., 1999. Phylogeny of Palaearctic Syrphidae (Diptera):
778 evidence from larval stages. *Zoological Journal of the Linnean*
779 *Society*, 127(1), pp.1-112.
- 780 Rotheray, G. and Gilbert, F., 2008. Phylogenetic relationships and the larval head
781 of the lower Cyclorrhapha (Diptera). *Zoological Journal of the Linnean*
782 *Society*, 153(2), pp.287-323.
- 783 Rotheray, G.F. and Gilbert, F., 2011. *The Natural History of Hoverflies*. Forrest
784 Text, Cardigan, UK.
- 785 Ruxton, G.D., Sherratt, T.N., Speed, M.P., Speed, M.P. and Speed, M.,
786 2018. *Avoiding Attack: The Evolutionary Ecology of Crypsis, Warning*
787 *Signals and Mimicry*. Oxford University Press.
- 788 Sherratt, T.N., 2002. The evolution of imperfect mimicry. *Behavioral*
789 *Ecology*, 13(6), pp.821-826.
- 790 Sherratt, T.N. and Peet-Paré, C.A., 2017. The perfection of mimicry: an
791 information approach. *Philosophical Transactions of the Royal Society B:*
792 *Biological Sciences*, 372(1724), p.20160340.
- 793 Speed, M.P. and Ruxton, G.D., 2010. Imperfect Batesian mimicry and the
794 conspicuousness costs of mimetic resemblance. *The American Naturalist*,
795 176(1), pp.1-14.
- 796 Speight, M.C.D. 2018. Species accounts of European Syrphidae. Syrph the Net
797 Database of European Syrphidae (Diptera) 103: 1-305

- 798 Speight, M.C.D. and de Courcy Williams, M., 2018. European Syrphid Genera
799 2018: Portraits of representative species. *Syrph the Net: The Database of*
800 *European Syrphidae*. Volume 102. Syrph the Net publications, Dublin.
- 801 Ståhls, G., Hippa, H., Rotheray, G., Muona, J. and Gilbert, F., 2003. Phylogeny
802 of Syrphidae (Diptera) inferred from combined analysis of molecular and
803 morphological characters. *Systematic Entomology*, 28(4), pp.433-450.
- 804 Stubbs, A.E. and Falk, S.J., 2002. *British Hoverflies: An Illustrated Identification*
805 *Guide*. British Entomological and Natural History Society.
- 806 Taylor, C.H., Gilbert, F. and Reader, T., 2013. Distance transform: a tool for the
807 study of animal colour patterns. *Methods in Ecology and Evolution*, 4(8),
808 pp.771-781.
- 809 Taylor, C.H., Reader, T. and Gilbert, F., 2016a. Why many Batesian mimics are
810 inaccurate: evidence from hoverfly colour patterns. *Proceedings of the*
811 *Royal Society of London B: Biological Sciences*, 283(1842), p.20161585.
- 812 Taylor, C.H., Reader, T. and Gilbert, F., 2016b. Hoverflies are imperfect mimics
813 of wasp colouration. *Evolutionary Ecology*, 30(3), pp. 567-581.
- 814 Taylor, C.H., Warrin, J., Gilbert, F. and Reader, T., 2017. Which traits do
815 observers use to distinguish Batesian mimics from their models?
816 *Behavioral Ecology*, 28(2), pp.460-470.
- 817 Tryjanowski, P., Pawlikowski, T., Pawlikowski, K., Banaszak-Cibicka, W. and
818 Sparks, T.H., 2010. Does climate influence phenological trends in social
819 wasps (Hymenoptera: Vespinae) in Poland? *European Journal of*
820 *Entomology*, 107(2), pp.203-208.

821 Wilson, J.S., Jahner, J.P., Williams, K.A. and Forister, M.L., 2013. Ecological and
822 evolutionary processes drive the origin and maintenance of imperfect
823 mimicry. *PloS One*, 8(4), p.e61610.

824 Young, A.D., Lemmon, A.R., Skevington, J.H., Mengual, X., Ståhls, G., Reemer,
825 *et al.*, 2016. Anchored enrichment dataset for true flies (order Diptera)
826 reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC*
827 *Evolutionary Biology*, 16(1), p.143.

Figures and Tables

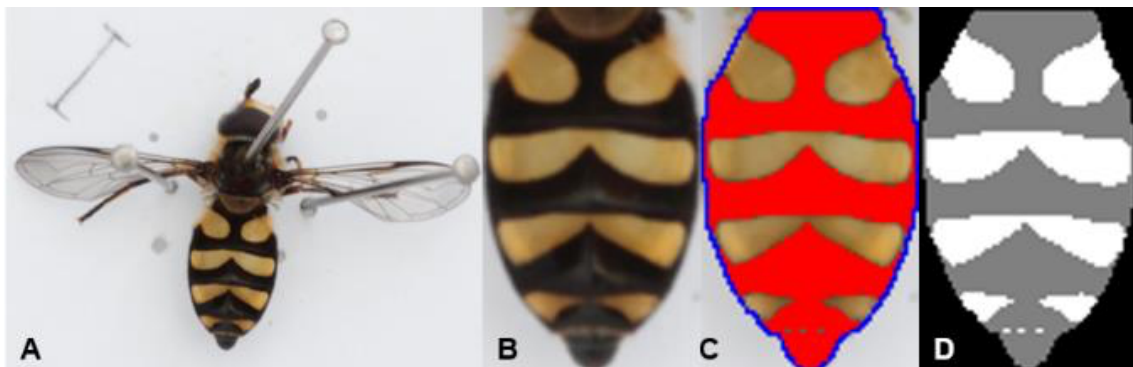


Figure 1 – The stages of image pre-processing: A) Original image of *Didea fasciata*. B) After rotation, cropping and scaling. C) Abdomen outlined in blue and black areas masked with red using ImageJ. D) Final binary image from MATLAB.

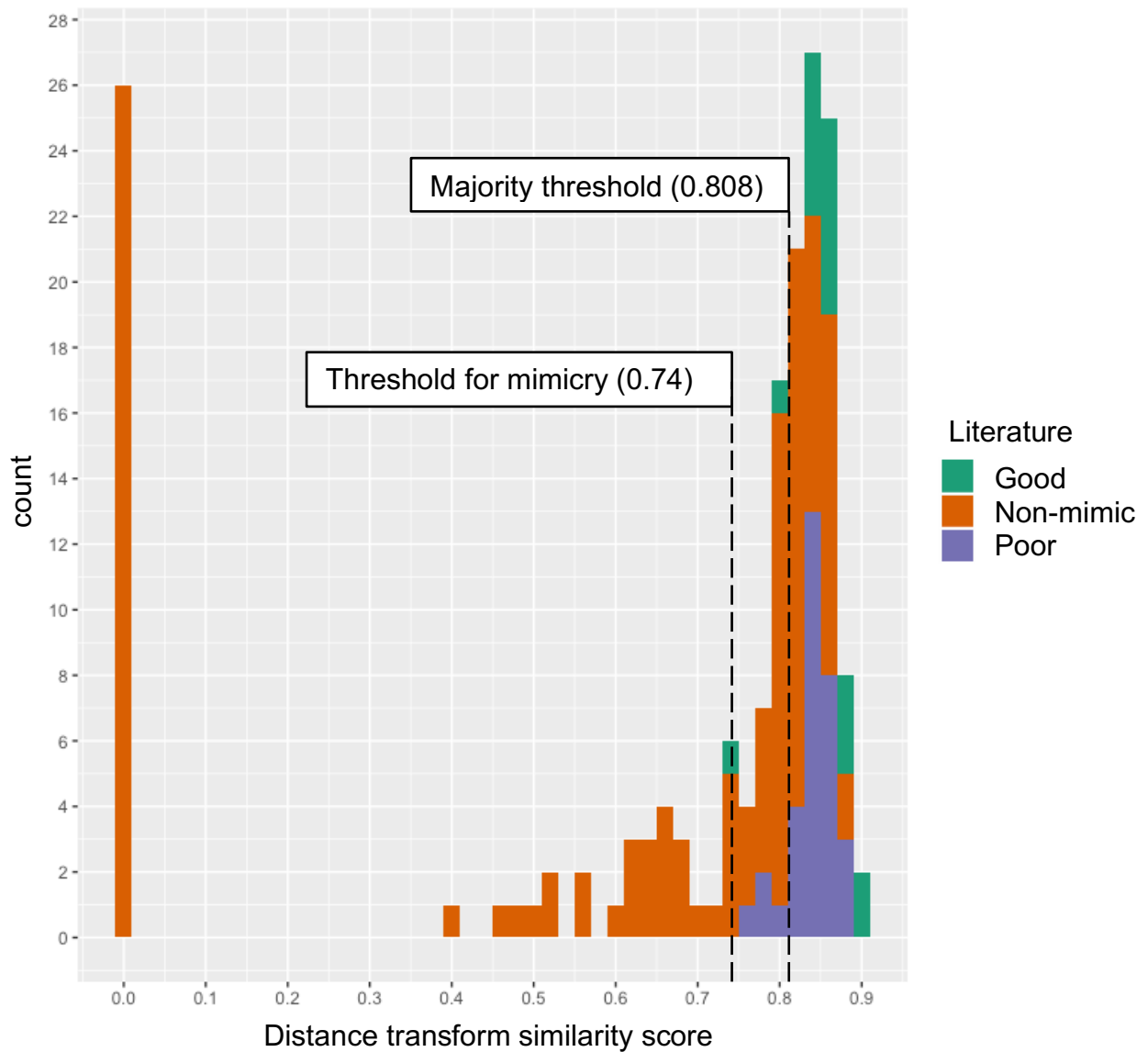


Figure 2 – Frequency distribution of similarity scores describing the accuracy of social wasp mimicry in 167 species of hoverfly, colour coded according to categories identified from the literature. The threshold for mimicry divides possible mimics from species that have never been considered to be mimics by experts, while the majority threshold marks the point above which most species are considered mimics. Bin width = 0.02.

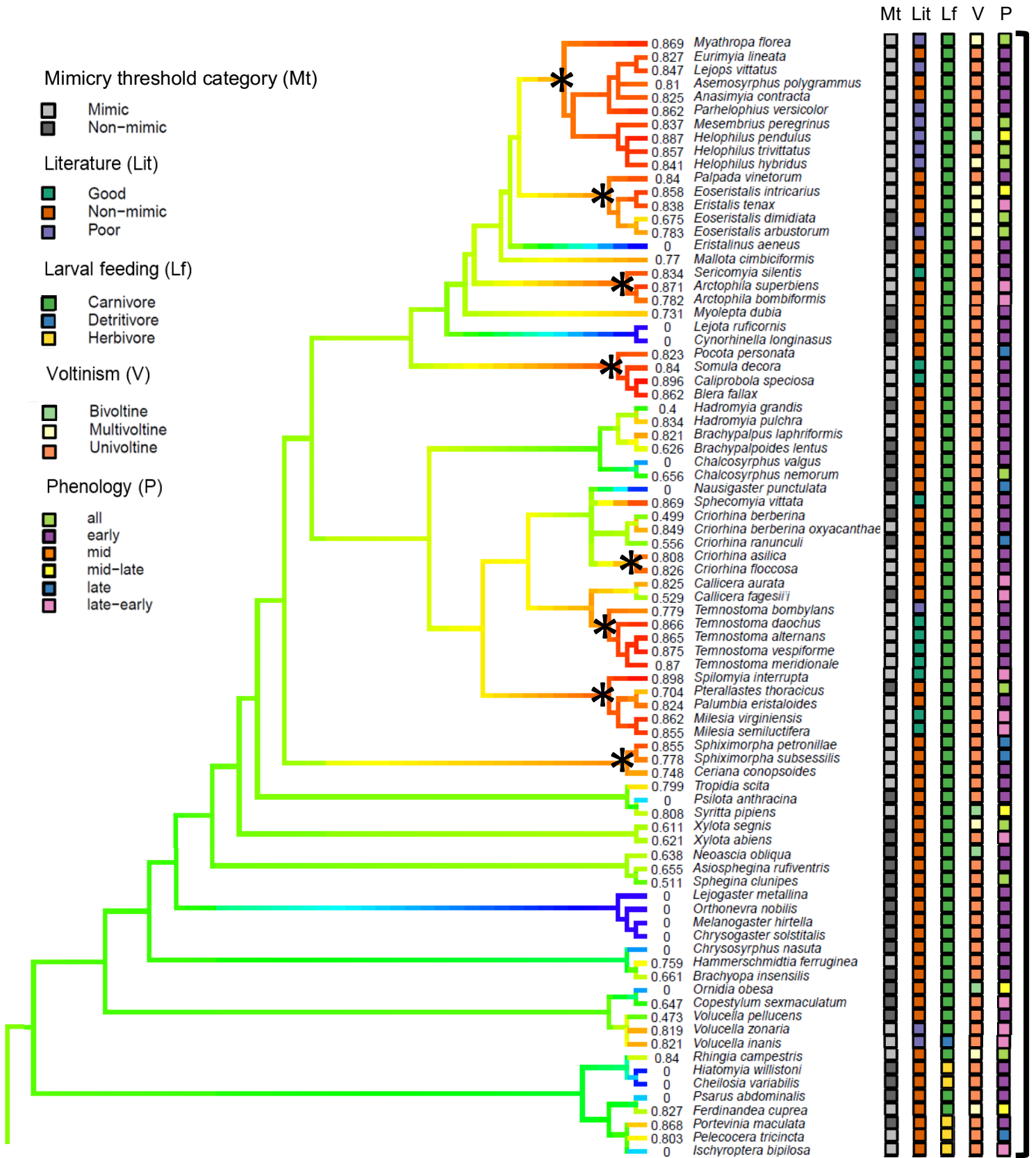
828 **Table 1** - A summary of the major conclusions obtained from the main *Vespula germanica* analysis, and whether they were
 829 supported by our different supplementary analyses and sensitivity tests. Each box refers to evidence either in support of (in
 830 bold) or in contrast to each conclusion. 'NA' means this conclusion was not tested in this analysis. *not including species with
 831 all-black abdomens.

Conclusion	Supplementary analyses			Sensitivity tests		
	Human survey	Binary analysis – mimicry threshold	Binary analysis - majority threshold	Wasp model type	Hairiness	PGLS with one species per genus
Identity of the top three and bottom three mimic taxa*	Supplementary dataset	NA	NA	Figure S2	Figure S2	NA
Location of threshold to divide mimics and non-mimics	NA	NA – threshold used in analysis	Figure 2	Figure S13	Figure S4	NA
PGLS was a better fit than OLS	Table S3	NA	NA	Table S3	Table S3	NA
Mimicry has evolved many times	NA	Figure 3	Figure S11	Figure S13	Figure S12	NA
Mimicry is sometimes lost in clades of good mimics	NA	Figure 3	Figure S11	Figure S13	Figure S12	NA
Phylogenetic signal for wasp mimicry is significant but not strong	Section 3.4	Table S6	Table S6	Figure S13	Figure S12	Table S6 – weak signal

Mapping the evolution of accurate mimicry

The best predictors of mimetic accuracy were wing length and voltinism	Table S3	Table S5	Table S5 – only wing length	Table S3	Table S3	Table S5
Smaller species are significantly less mimetic than larger species	Table 2	Table S6	Table S6	Table S4	Table S4	Table S6
Univoltine species are significantly less mimetic than multivoltine species	Table 2	Table S6	NA	Table S4	Table S4	Table S6

Mapping the evolution of accurate mimicry



Mapping the evolution of accurate mimicry

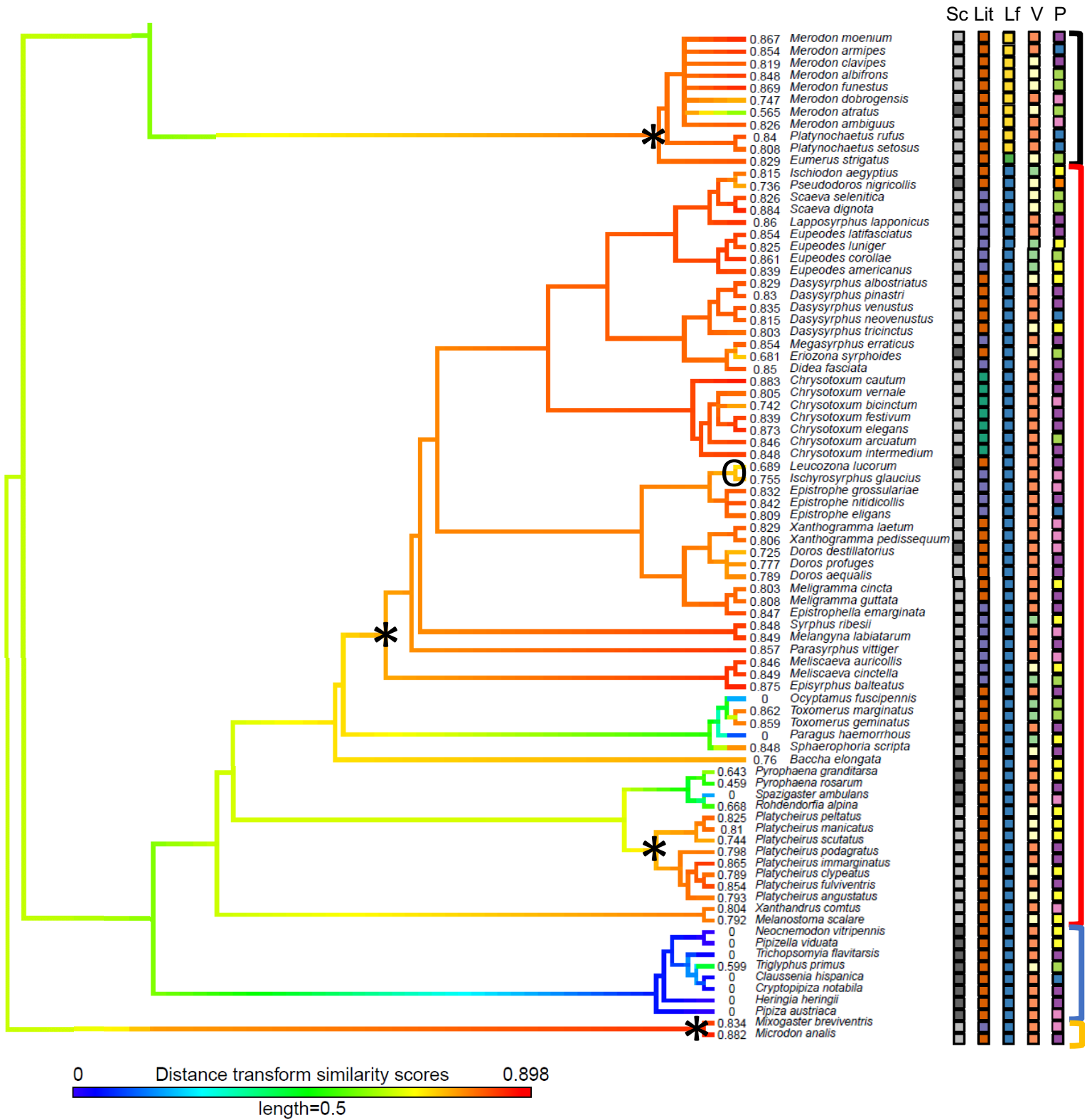


Figure 3 – A literature-derived phylogeny of 167 hoverfly species. Warmer tip colours represent higher similarity to, and hence better mimicry of, the social wasp *V. germanica*. Tips are labelled with the distance transform similarity scores and a colour-coded grid to represent the ecological traits investigated (for abbreviations, see key). Defining mimicry as a binary trait using the mimicry threshold (0.74) allowed us to identify ancestral nodes where social wasp mimicry evolved (*****) and was lost (**O**) according to ‘fastAnc’ ancestral state estimates under Brownian evolution. Blank nodes before a ***** are non-mimetic. Subfamilies (indicated by the brackets on the far right): black = Eristalinae, red = Syrphinae, blue = Pipizinae, yellow = Microdontinae (Chandler, 1998; Stubbs and Falk, 2002).

Table 2 – Coefficients from the best PGLS models describing the relationship between life history traits and mimetic similarity scores for 167 species of hoverfly for *Vespula germanica*. Similarity scores were either calculated by pattern analysis (“distance transform”) or from a survey of human volunteers (“survey”). SEM – standard error.

		Coefficients	SEM	t-value	p-value	
Distance transform	Intercept (Univoltine)	0.284	0.159	1.783	<0.001	
	Wing length	0.041	0.010	4.074	<0.001	
	Voltinism	Bivoltine	0.116	0.051	2.293	0.023
		Multivoltine	0.163	0.072	2.263	0.025
Survey	Intercept (Univoltine)	1.696	0.590	2.875	0.005	
	Wing length	0.122	0.049	2.467	0.015	
	Voltinism	Bivoltine	-0.116	0.269	-0.430	0.668
		Multivoltine	1.022	0.384	2.661	0.009