doi:10.1111/evo.14392

Estimating hybridization in the wild using citizen science data: A path forward

Nicholas R. Minor,[1,2,*] Paul J. Dougherty,[1,2,3,*] Scott A. Taylor,[4] and Matthew D. Carling[1,2,3,5]

[1] *Department of Zoology & Physiology, University of Wyoming, Laramie, Wyoming 82071*

[2] *Program in Ecology, University of Wyoming, Laramie, Wyoming 82071*

[3] *University of Wyoming Museum of Vertebrates, Laramie, Wyoming 82071*

[4] *Department of Ecology and Evolutionary Biology, University of Colorado–Boulder, Boulder, Colorado 80309*

[5] *E-mail: mcarling@uwyo.edu*

Received February 24, 2021
Accepted September 19, 2021

Genomic evidence of introgression in natural populations has reinvigorated the study of hybridization in recent years. Still, it is largely unknown how frequently individual organisms mate across species lines. Recently, Justyn et al. suggested that eBird, one of the world's largest citizen science databases, may supply adequate data for estimating hybridization rates. Here, we compare Justyn et al.'s estimates—and their conclusions that hybridization is rare—with estimates from museum and molecular data. We also estimate hybridization using eBird observations from areas and times when hybridization is possible, namely, in contact zones during the breeding season. These estimates are all considerably higher than those reported in Justyn et al., emphasizing that inferences from multiple datasets can differ radically. Finally, we demonstrate an approach for predicting the location of hybrid zones using eBird data, which can be done with high confidence and with unprecedented resolution. We show that citizen science data, far from settling the question of how frequently bird species hybridize, instead offer a promising step toward more focused study of hybrid zones.

KEY WORDS: Citizen science, eBird, hybridization, introgression.

Interest in hybridization has exploded in recent years, largely thanks to a burgeoning number of examples of interspecific introgression (Seehausen 2004; Mallet 2005; Taylor and Larson 2019). These examples have led many researchers to posit that hybridization is ubiquitous across the tree of life, explaining patterns like gene tree-species trees discordance (Degnan and Rosenberg 2009; Ottenburghs et al. 2017; Drovetski et al. 2018) or reticulate phylogenies (Funk 1985; Edwards 2009; Mallet et al. 2016; Everson et al. 2019). Still, it is largely unknown how often individual organisms hybridize. For birds, the single published estimate prior to 2020 comes from Mayr (1963), where Mayr observed approximately one hybrid for every 60,000 examined bird study skins (Mayr 1963). Recently, Justyn et al. 2020 suggested that eBird (http://www.ebird.org), a citizen science project where data are submitted by amateur volunteers, is a solution to this impasse (Sullivan et al. 2014; Justyn et al. 2020). Justyn et al. argued that the massive scale of the eBird database, which contains more than 600 million observations, may capture enough cases of hybridization to make accurate estimates (Johnston et al. 2020).

Although this is an exciting possibility, others have identified a number of downsides in using eBird data for this purpose. Ottenburghs and Slager (2020), for example, pointed out that citizen scientists who contribute to eBird are likely to underreport many hybrids (Ottenburghs and Slager 2020). Justen et al. (2020) extended this argument further, showing that common yet difficult-to-identify hybrids are reported far less than the occasional spectacular hybrid, despite the fact that many hybrids are likely to be cryptic (Coyne et al. 2007; Hewitt 2008; Vallender et al. 2009; Campagna et al. 2017; Justen et al. 2020). This is especially the case if citizen scientists are inexperienced with the technicalities of bird identification. Justen et al. (2020) also

[*] Nicholas R. Minor and Paul J. Dougherty should be considered joint first authors.

argued that hybridization rates can only reasonably be calculated using observations in areas where two species have the opportunity to hybridize. By contrast, Justyn et al. (2020) included observations from locations where two species' ranges do not overlap, which artifactually deflates hybridization estimates.

Thus far, estimating hybridization rates using eBird data has been descriptive, calculating the proportion of hybrids to parentals. This approach has many benefits; it is straightforward, even with an immense dataset like eBird, and it uses all available eBird observations. However, the challenge of how to filter the eBird dataset remains.

One way to avoid the downsides of using eBird data to estimate hybrid prevalence is to use eBird data to *predict* the presence of hybrids instead of describing hybridization as a proportion. To demonstrate the efficacy of these predictions, we applied eBird and museum data to accomplish three aims. First, we attempted to explicitly replicate Justyn et al.'s (2020) analysis and compare those results with analyses based on a geographically constrained eBird dataset and a museum dataset. Second, we compared eBird-based estimates of hybridization with estimates based on molecular data and museum specimens. And third, we predicted relative abundance of *Passerina cyanea, Passerina amoena*, and their hybrids across North America. We did so by modeling eBird observation count as a function of latitude, longitude, elevation, habitat classification, bioclimatic variables, as well as eBird effort covariates, including day of year, observation start time, checklist duration (in minutes), distance traveled (in kilometers), number of observers, and observation type. The resulting distribution maps allow us to identify areas of breeding range overlap where hybridization may occur.

Critically, our approach avoids the conundrum of describing hybrid proportions where there are no, or very few, eBird submissions. In such cases, the presence of hybrids—including rarely identified cryptic hybrids—can be predicted with uncertainty, providing citizen scientists new avenues to participate in science by ground-truthing predictions. Citizen scientists could thus be incentivized to fill gaps in the eBird database, advance their identification skills, and ultimately deepen their engagement with large-scale science and the natural world—all in line with the mission of the eBird enterprise (Sullivan et al. 2014).

Estimating Hybrid Proportions with eBird Data

To compare with previous estimates, we used the package *auk* to subset the full eBird dataset (version US_relJan2020) to observations from 2010 through 2018 for the entire United States, as per Justyn et al. 2020 (Strimas-Mackey et al. 2018; Justyn et al. 2020). Importantly, the package reduces redundancy from the dataset by collapsing all observations of the same individual bird into single observations. We then calculated the proportion of hybrid individuals to nonhybrid parentals to compare directly with the eBird-based estimates in Justyn et al. 2020 (all scripts used in the processing of these data can be found at https://github.com/mcarling/avianhybridization).

Unfortunately, we were unable to replicate the rangewide hybrid proportions reported in Justyn et al. 2020, both for a subset of the hybrids examined in that paper, as well as for *P. cyanea* and *P. amoena* specifically (Table 1). This discrepancy may stem from differences in our filtering procedure, which left us with 1,179,687 individual hybrids out of 3,668,813,292 total counts of individuals—significantly more than the 212,875 hybrids and 334,770,194 observations used by Justyn et al. 2020. Justyn et al.'s numbers may represent *presences* of each species, where, as an example, six *P. amoena* are equal to one observation of that species, as opposed to our procedure, which retains six *P. amoena* as six individual observations.

Nevertheless, rangewide estimates of hybrid proportions, where the denominator includes parental individuals from year round, may not reflect hybridization rates in any one locality (Justen et al. 2020). As such, we filtered the eBird dataset further to contain only those observations made during the majority of species' breeding seasons, which we considered as June and July, though in many locations, mid-May to mid-July may be more representative. To eliminate individuals observed outside areas where hybridization is likely to occur, we further pruned the dataset to observations made in counties where both parental species were observed, or where hybrids and one or the other parentals were observed. Although it is possible that hybrids are found outside a hybrid zone, we argue that the majority of hybrids are found in hybrid zones. This is the case in the eBird dataset, where the proportion of birds identified as hybrids is highest in areas where both parental species occur. Next, we calculated hybrid proportions for 11 species pairs that are well known to hybridize and are likely to be represented by eBird observations, including Baltimore and Bullock's Oriole, Western and Glaucous-winged Gull, Mallard and Mottled Duck, and Tufted and Black-crested Titmouse (see Supporting Information Table S1). Species-specific hybrid proportions are considerably higher, often by as much as two orders of magnitude, when based only on observations from locations where and times when hybridization is possible.

Estimating Hybrid Proportions with Museum Data

Hybrid proportions based on observations at times and locations where hybridization is possible may more accurately reflect reality. However, it is possible that this procedure still underestimates the true extent of hybridization. Many hybrid species are

Table 1. Comparison of hybrid proportion estimates based on eBird data.

	Rangewide, year round hybrid proportions from Justyn et al. (2020) (%)	Rangewide, year round hybrid proportion from this study (%)	Rangewide, breeding season estimates (%)	Hybrid zone, breeding season estimates (%)
Full eBird dataset	0.064	0.032	0.036	NA
eBird dataset, no monotypic genera	NA	0.036	0.039	NA
Passerina cyanea	0.062	NA	0.029	4.728
Passerina amoena	0.392	NA	0.218	0.290

In this table, we calculate the percentage of hybrid observations out of all observations in the eBird dataset from between 2010 and 2018 in North America and compare with the same estimates in Justyn et al. (2020). We also compare our estimates of *Passerina cyanea* and *P. amoena*. Additionally, we demonstrate how the estimates of hybrid frequency increase when they are based only on observations from within the contact zone between those two species during the breeding season.

Table 2. The influence of dataset filtering and data source on hybrid proportions for *P. cyanea* and *P. amoena*.

	Passerina amoena	*Passerina cyanea*
Year round, rangewide hybrid proportions based on museum data (Justyn et al. 2020) (%)	1.25	0.48
Year round, rangewide hybrid proportions based on museum data, this study (%)	3.29	1.58
June–July, rangewide hybrid proportions based on museum data (%)	5.11	4.04
June–July, hybrid zone hybrid proportions based on museum data (%)	6.92	19.39

Both the data filtering procedure and the source of data have considerable influence on hybrid percentages. Hybrid zone and breeding season estimates based on museum data, especially for *Passerina cyanea*, are orders of magnitude larger than those based on rangewide, year-round museum data reported in Justyn et al. (2020).

difficult to identify visually (e.g., Manthey and Robbins 2016; Linck et al. 2019), especially when technical expertise in hybrid identification is uncommon among citizen scientists.

To explore this possibility, we repeated our hybrid proportion calculations for *P. cyanea* and *P. amoena* with museum specimen data, which we accessed via VertNet (http://www.vertnet.org). Museum specimens allow more careful measurements and even molecular analysis, which increases the likelihood that museum workers will detect highly cryptic hybrids compared to visual observers. Furthermore, museum specimen data are available in locations within the *P. cyanea* x *P. amoena* hybrid zone with few to no observations in the eBird dataset. Although this better coverage may confound comparisons with eBird-based proportions, museum-based proportions may better represent this well-known hybrid zone.

Like with the eBird data, we calculated rangewide/year-round hybrid proportions, rangewide/breeding season proportions, and hybrid zone/breeding season proportions. These museum-based proportions, including the rangewide/year-round baseline (or lower bound, as suggested in Ottenburghs and Slager 2020), were considerably higher than our eBird-based proportions, and especially higher than proportions reported in Justyn et al. (2020) (Table 2). Because these museum specimens provide more lines of evidence to detect and confirm hybrids, we suggest that museum-based proportions represent the upper bound to match the lower bound provided by eBird-based proportions, at least for the *P. cyanea* x *P. amoena* species pair.

Comparing eBird-based and Molecular Estimates of Hybrid Prevalence in Sphyrapicus

Museum data may be adequate to represent hybrid frequencies in some hybrid zones, but in others, hybrids may be so difficult to identify visually that they are even missed by experts. Recent molecular work has described frequent hybridization between visually near-identical parental taxa (e.g., *Contopus* wood-pewees), hybrids between which would be impossible to identify based on intermediate features or combinations of plumage characteristics (Manthey and Robbins 2016). Even in systems where parental taxa have obvious plumage differences, the relationship between ancestry and appearance may not be straightforward.

Table 3. **eBird-based hybrid proportions compared with molecular hybrid proportions from the hybrid zone between *Sphyrapicus ruber* and *S. nuchalis*.**

County, State	eBird hybrid percentage (n)	Molecular hybrid percentage (n)
Lassen, California	0 (0 hybrids, 110 total)	20 (2 hybrids, 10 total)
Modoc, California	4.76 (7 hybrids, 147 total)	100 (30 hybrids, 30 total)
Trinity, California	0 (0 hybrids, 81 total)	10 (10 total, 1 hybrid, 10 total)
Crook, Oregon	11.53 (3 hybrids, 26 total)	22.22 (2 hybrids, 9 total)
Grant, Oregon	0 (0 hybrids, 43 total)	71.42 (5 hybrids, 7 total)
Jackson, Oregon	0 (0 hybrids, 496 total)	0 (0 hybrids, 3 total)
Josephine, Oregon	0 (0 hybrids, 127 total)	0 (0 hybrids, 8 total)
Klamath, Oregon	0.43 (2 hybrids, 463 total)	0 (0 hybrids, 2 total)
Lake, Oregon	8.28 (29 hybrids, 350 total)	100 (50 hybrids, 50 total)
Wallowa, Oregon	0 (0 hybrids, 28 total)	0 (0 hybrids, 1 total)
Wheeler, Oregon	0 (0 hybrids, 11 total)	20 (1 hybrid, 5 total)
Yakima, Washington	1.68 (11 hybrids, 651 total)	0 (0 hybrids, 1 total)

Sample sizes include all parentals and hybrids sampled. Focal counties were those sampled for genomic analysis in Billerman et al. (2019), which found that the prevalence of genomically admixed individuals within the *Sphyrapicus* hybrid zone far exceeded expectations based on visual discrimination alone. In Oregon's Lake county and California's Modoc county, for example, 100% of individuals subject to molecular analysis were admixed.

Instead, many hybrids, especially cryptically colored females, may appear indistinguishable from either parental species (Baiz et al. 2020; Thompson et al. 2020). Targeted collection of data, where identifying hybrids does not depend on visual discernment, may be the only solution to these problems.

For these reasons, we brought previously published molecular data to bear as a third source of hybrid data, this time from the hybrid zone between Red-naped (*Sphyrapicus nuchalis*) and Red-breasted Sapsuckers (*Sphyrapicus ruber*). Billerman et al. (2019) showed that individuals with hybrid ancestry (genome-wide hybrid index between 0.1 and 0.9) are often cryptic enough that observers may be unable to correctly identify the vast majority of hybrid individuals (Billerman et al. 2019). We compared our eBird-based proportions with proportions based on whole-genome data within those counties where the genome data were collected. Again, we find that eBird data vastly underestimate the frequency of hybrids in the *Sphyrapicus* hybrid zone (Table 3, Supporting Information Fig. S5).

Additionally, we emphasize that far from being binary, hybridization occurs along a spectrum, where many admixed individuals beyond the F1 generation may be undetectable without molecular data. As such, for the *Sphyrapicus* hybrid zone, we suggest that molecular data provide the best available upper bound for estimates of hybrid frequency for two reasons. First, molecular data allow researchers to identify admixed individuals even when they are not visually discernible. Second, molecular data make it possible to identify higher generational hybrids and minorly admixed individuals, enabling a more liberal definition of hybrids than perhaps Justyn et al. (2020) and others may have intended to address.

Predicting Areas of Overlap between Hybridizing Parental Species

The eBird dataset contains vastly more observations of *P. cyanea* and *P. amoena* than it does for hybrids between the two. Extensive parental observations make possible the first step in our predictive approach: to demonstrate how eBird data can be used to estimate where the two species overlap and have the opportunity to hybridize. These predicted areas of overlap present the first opportunity for ground-truthing by citizen scientists.

To determine areas where hybridization between *P. amoena* and *P. cyanea* is possible, we predicted the expected number of individuals of each parental species that the average eBird observer could be expected to record—what we call relative abundance—across all 2.5 km × 2.5 km cells in North America (Strimas-Mackey et al. 2018). After downloading all eBird records for *P. cyanea* and *P. amoena* as well as sampling event data from the eBird basic dataset (version ebd_relMar-2020), we used the *auk* R package to filter these datasets down to records from stationary and traveling checklists submitted in the United States, Canada, and Mexico between June 1 and July 16 from 2010 to 2018. We excluded records submitted after July 16 because many *P. amoena* initiate fall migration in late July (Young 1991; Rohwer et al. 2005; Pyle et al. 2009).

We further subset the eBird datasets to include only observations from checklists with fewer than 10 observers, shorter than 5 hours, and on which observers traveled fewer than 5 km (Strimas-Mackey et al. 2018). We then merged the sampling event dataset with eBird records to identify the checklists on which *P. cyanea* were recorded, and then filled the dataset with zeroes for the remaining locations where *P. cyanea* was not observed. We then used the same process to create datasets for *P. amoena* and for hybrids. For all checklists in these "zero-filled" datasets, we calculated the proportion of 16 landcover types (MODIS MCD12Q1 v006 product; Friedl and Sulla-Menashe 2015; Strimas-Mackey et al. 2020), elevation median and standard deviation (EarthEnv 1 km resolution elevation data from GMTED2010 product; Amatulli et al. 2018), and values for 19 bioclimatic variables (WordClim database; Hijmans et al. 2005) in 2.5 km × 2.5 km cells centered on the provided latitude and longitude.

To predict the relative abundances of *P. cyanea* and *P. amoena* across North America, we first fit a generalized additive model for the zero-filled eBird datasets for each species using the mgcv package in R (Wood and Wood 2015). For both species, we defined checklist observation count as a function of year day, checklist duration, checklist distance, the number of observers, the time of day checklist observations started, checklist protocol type (stationary or travelling), latitude, longitude, elevation median, a subset of land cover classification covariates (water, evergreen needleleaf forest, deciduous broadleaf forest, closed shrubland, open shrubland, woody savanna, grassland, cropland, urban, and built-up land), and a subset of bioclimatic variables (temperature during the warmest quarter, annual precipitation, and precipitation during the warmest quarter). We selected these three bioclimatic variables based on Carling and Thomassen (2012), who found them to be significant predictors of hybrid index (Carling and Thomassen 2012). We note that future researchers should consider the breeding biology of focal taxa when selecting land cover and bioclimatic variables for models of relative abundance. In these models, we defined five knots (the number of connection points between different model segments in the GAM) for each continuous predictor variable and a cyclic cubic spline with seven knots for time of day, as recommended by Keele (2007) and Strimas-Mackey et al. (2020). We also used cross-validation to verify that the specified number of knots optimizes the smoothing parameter in our models.

After fitting the models, we used the model output to predict relative abundance of each species across North America, which we used to produce heat maps of each species' breeding distribution (Fig. 1). Finally, we produced maps of overlap between the two species based on threshold relative abundances of 0.01, 0.05, 0.25, and 0.1 for both species (Supporting Information Figs. S1–S4, respectively). These maps show areas of potential hybridization with unprecedented detail and do so while also allowing us to quantify uncertainty. These advantages are thanks to the large number of observations (n = 491,520) for both *P. cyanea* and *P. amoena* with which to fit our relative abundance model.

Depending on the relative abundance threshold for parental species, our maps of parental species overlap coincide with previous estimates of the species pair's hybrid zone (Emlen et al. 1975; Carling and Zuckerberg 2011). However, these eBird-based estimates resolve the boundaries of the hybrid zone for the first time, to our knowledge, while also allowing users of this predictive approach to measure and map uncertainty. These estimates showcase the benefits of eBird data for predicting areas of overlap, an endeavor that may prove fruitful for other avian hybrid zones for which parental observations are abundant.

Predicting Hybrid Relative Abundance Using eBird Data

There are few observations of *P. cyanea* × *P. amoena* hybrids in our filtered eBird dataset ($n = 93$). However, it is still possible to apply the same relative abundance modeling approach, described above, using hybrid observations. We predicted the relative abundance of *P. cyanea* × *P. amoena* hybrids across North America by modeling eBird observation count as a function of latitude, longitude, elevation, habitat classification, bioclimatic variables, and effort covariates.

Using this approach, we mapped hybrid relative abundance (Fig. 2). This map shows hybrids to have extremely low relative abundance, often much less than 0.01. Surprisingly, our model projected hybrid relative abundance outside the typically acknowledged hybrid zone, though again at very low relative abundances. Although these predictions are intriguing, they are unlikely to reflect the realities of hybrid abundance and distribution. Compared to the small size of our predicted relative abundances, standard errors are immense, consistently coming out to between 1.7 and 7.89 across the hybrid zone. Standard errors were even higher in areas outside the hybrid zone, reaching values as high as 27.06. From both a statistical and biological perspective, we argue that these maps should be taken more as a proof-of-concept and less as an accurate representation of reality. Hybrid observations in the eBird dataset are too sparse to permit the fitting of linear models.

Predicting Hybrid Distribution Using Bioclimatic Variables

Next, we attempted to predict the geographic distribution of hybrids using both eBird observations and museum records with bioclimatic variables. These predictions were based on the match of any given 2.5 × 2.5 km cell within North America to the bioclimatic conditions exhibited at each location a *P. cyanea* × *P. amoena* hybrid was observed. We used the BIOCLIM algorithm (Hijmans et al. 2005) to match conditions in each cell containing a hybrid observation and conditions in all other cells. The BIOCLIM algorithm produces a percentile distribution of the values

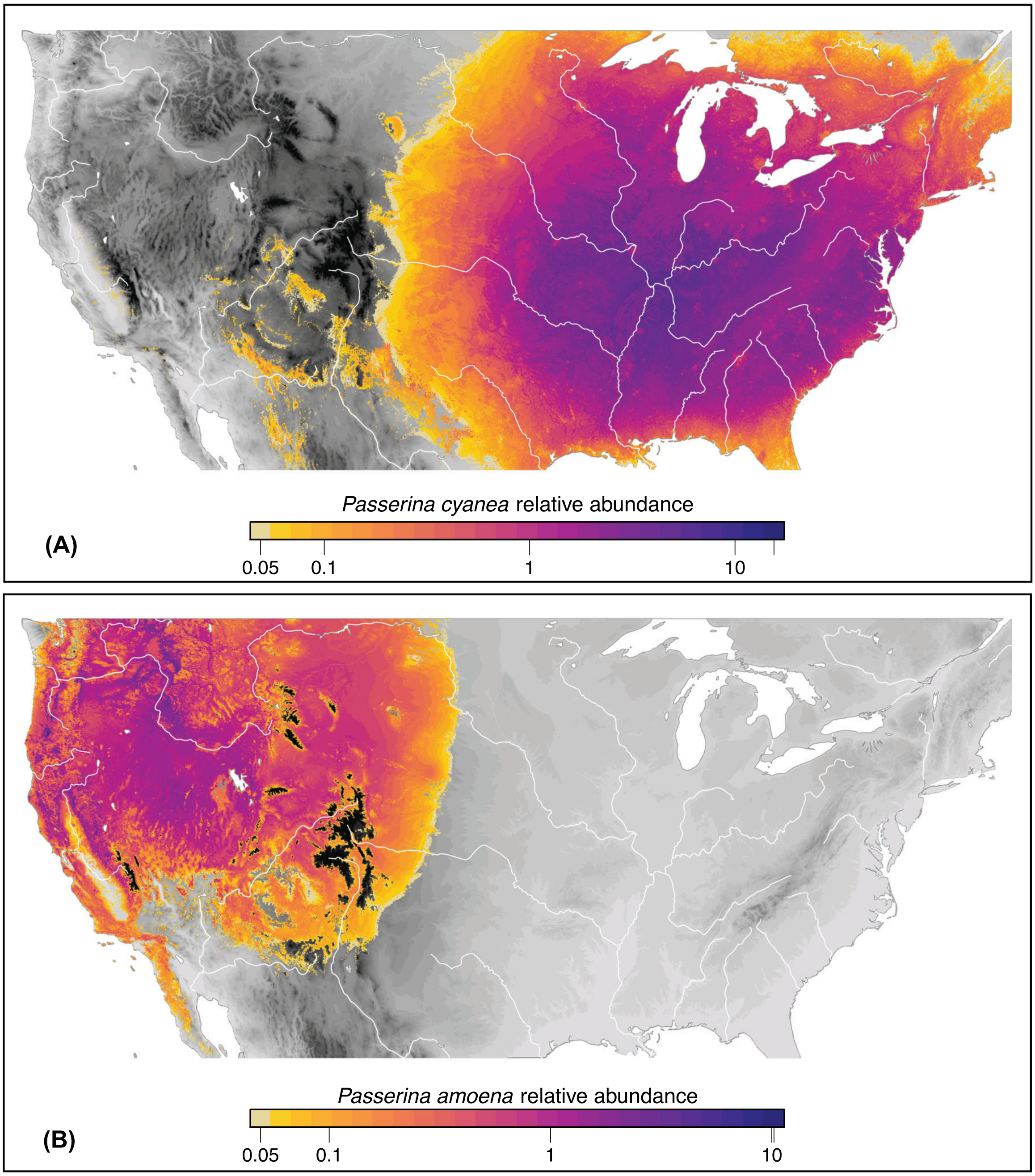
Passerina cyanea relative abundance
(A)
0.05
0.1
1
10
Passerina amoena relative abundance
(B)
0.05
0.1
1
10

Figure 1. Log transformed relative abundances of *P. cyanea* (A) and *P. amoena* (B).

at known locations of occurrence ("training sites"), and computes similarity of other locations based on their overlap with this percentile distribution. Locations are more suitable the closer they are to the 50th percentile.

The major strength of this approach, in contrast to our modeling of hybrid relative abundance, is that eBird observations can be used in tandem with museum records. In this case, our sample size was bolstered considerably ($n = 93$ for eBird alone, $n = 302$ for the combined dataset). This is because locations without known hybrid occurrences do not need to be zero-filled. To examine the efficacy of combining eBird and museum data, we produced bioclimatic distribution maps with museum data alone as well as with both datasets.

Although predicted suitable areas range far more widely when both eBird and museum data are included, both maps fail to recapitulate the known hybrid zone between *P. cyanea* and *P. amoena* (Fig. 3), as they are both more extensive than the known hybrid zone and also lack the southern portion. This may be due to overrepresentation of northern hybrids and underrepresentation of southern hybrids in both museums and eBird.

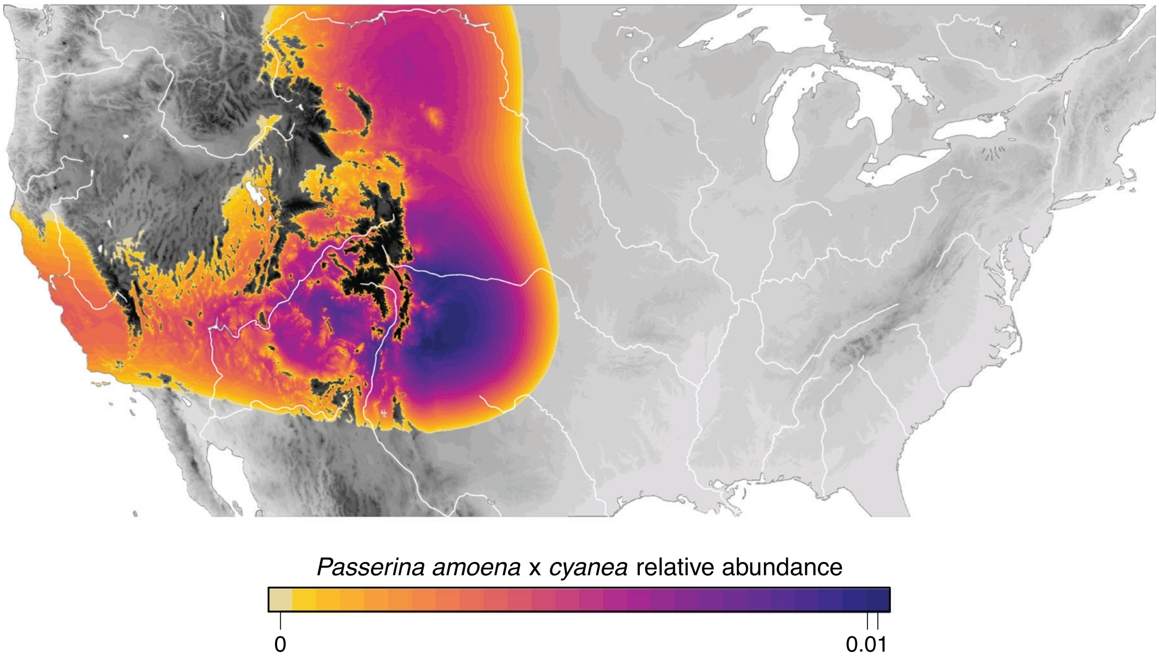

Figure 2. Log transformed Relative abundance of *Passerina cyanea* × *P. amoena* hybrids based on 93 available eBird observations from 2010 to 2018. Predicted hybrid relative abundances were quite low, ranging between 0 to 0.01. However, we suggest this is due to inadequate sample size with which to fit our relative abundance model. This is reflected in the high uncertainty, which was orders of magnitude larger than the relative abundance estimates themselves.

Predicting Parental Taxon Abundance Ratios within Hybrid Zones

Above, we demonstrate that it is possible to use eBird data to predict (1) the relative abundance of *P. cyanea* × *P. amoena* hybrids, and (2) the bioclimatic distribution of hybrids. However, in doing so we also demonstrate that the small number of hybrid observations in eBird limits the power of these predictions. In both predictions, the standard error was large enough to render relative abundance and bioclimatic distribution estimates practically unusable. Relative abundance estimates, for instance, ranged from 0 to 0.01, whereas their standard error ranged from 1.5 to 27.06. As pointed out by previous authors (Justen et al. 2020; Ottenburghs and Slager 2020), the paucity of hybrid observations may stem from higher frequencies of cryptic hybrids than are known. Additionally, in the case of this particular hybrid zone, eBird observer effort is chronically low. Should this remain true for years to come, the number of hybrid observations relative to the number of parentals is likely to remain low, which means that predicting hybrid abundance with confidence may not be a matter of waiting until there are enough observations. Moreover, the challenges of hybrid identification and low observer effort may continue to plague many hybrid zones across North America, including those between Eastern and Western Wood-Pewees, Hermit and Townsend's Warblers, Red-breasted and Red-naped Sapsuckers, and others.

As such, we urge strong caution in predicting the relative abundance of hybrids themselves based on eBird data. However, as we have demonstrated, the dataset is well suited to predicting parental taxon relative abundance as well as areas of overlap (hybrid zones), an approach that remains underutilized (but see Taylor et al. 2014). With these established strengths of the dataset in mind, we offer our final approach for understanding avian hybrid zones using eBird data, which we focus once more on *P. cyanea* and *P. amoena*. We again mapped the area of overlap between *P. cyanea* and *P. amoena*—that is, areas where both species have a relative abundance of at least 0.05—and then added an additional layer depicting the ratio of relative abundances of the two species: RA_Indigo and RA_Lazuli. We then plotted both species relative abundance across a longitudinal transect of the hybrid zone.

The parental species abundance ratio maps (Fig. 4) show how the abundance of both species—a proxy of heterospecific and conspecific mate availability, assuming an even sex ratio—varies across locations where they have the opportunity to hybridize. This particular hybrid zone spans an environmental gradient from wetter to drier habitats, leading many to suggest that *P. cyanea* and *P. amoena* have undergone niche divergence into wetter and drier habitats, respectively (Swenson 2006; Carling

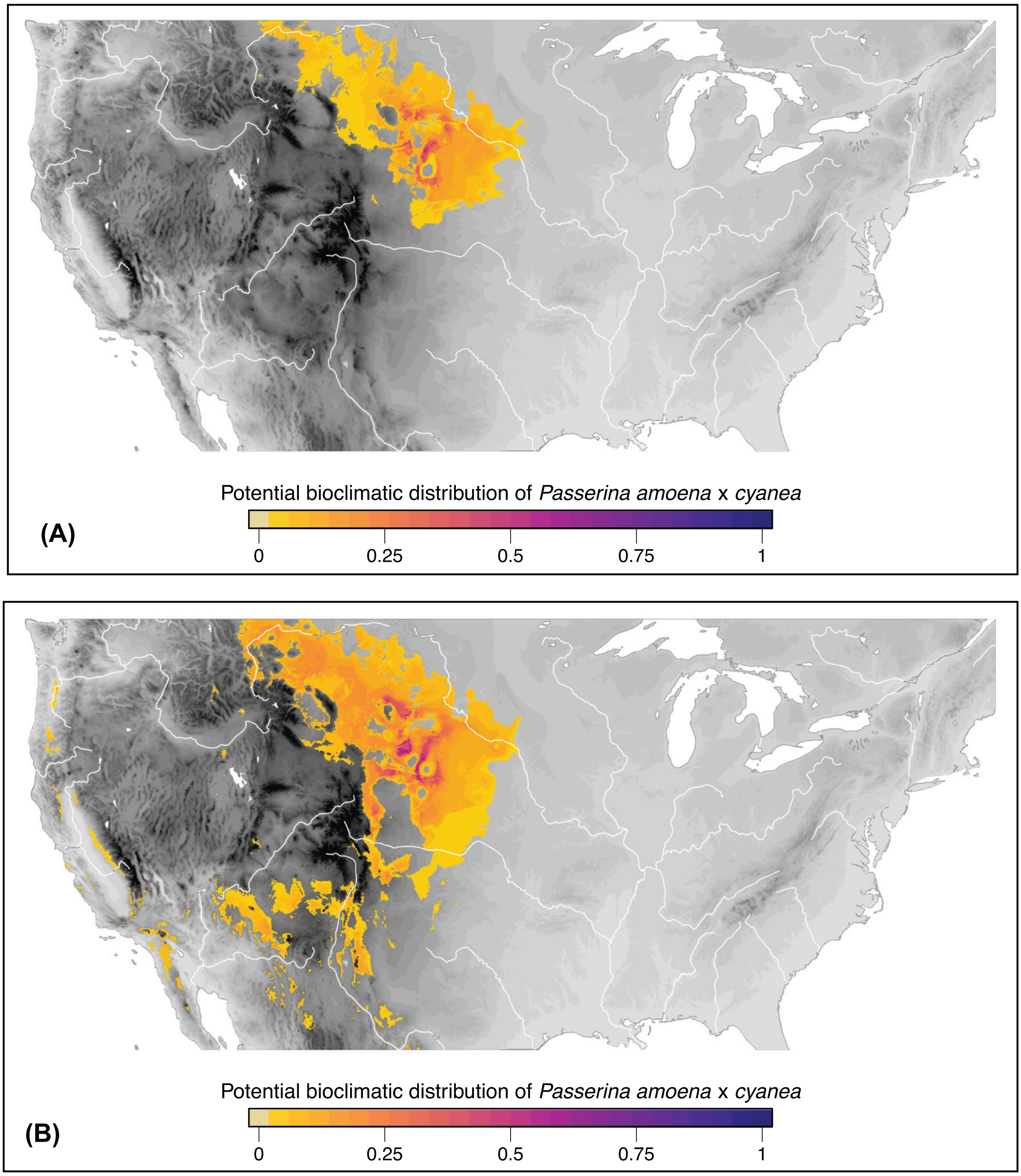
Potential bioclimatic distribution of *Passerina amoena* x *cyanea*
(A)
0
0.25
0.5
0.75
1
Potential bioclimatic distribution of *Passerina amoena* x *cyanea*
(B)
0
0.25
0.5
0.75
1

Figure 3. **Potential bioclimatic suitability based on (A) 209 museum records of *Passerina amoena* × cyanea hybrids; and (B) 89 available eBird observations (2010–2018) and 209 museum records (all time) of *Passerina amoena* × cyanea.**

and Thomassen 2012). Our plots of parental species relative abundance across longitude (Fig. 5 and Supporting Information Fig. S6) may show even deeper habitat segregation; both species are predicted to be notably less abundant in longitudes where the other species is present. This may be a signature of interspecific territoriality, which has previously been demonstrated between these two species (Emlen et al. 1975; Baker 1991), or of resource competition or maladaptation to the other species' preferred habitats.

Although these maps do not show hybridization rates or the expected frequency of hybrids, they do provide a clear foundation for exploring how parental species abundance, inferred with eBird data, relates to hybridization rates, which may be better inferred with complementary forms of data, such as encounter rates (Willis et al. 2011). We offer three hypotheses to be tested by future researchers integrating eBird data with other forms of hybridization data: (1) hybridization rates are uniform across the range of parental taxon abundances, implying that the availability

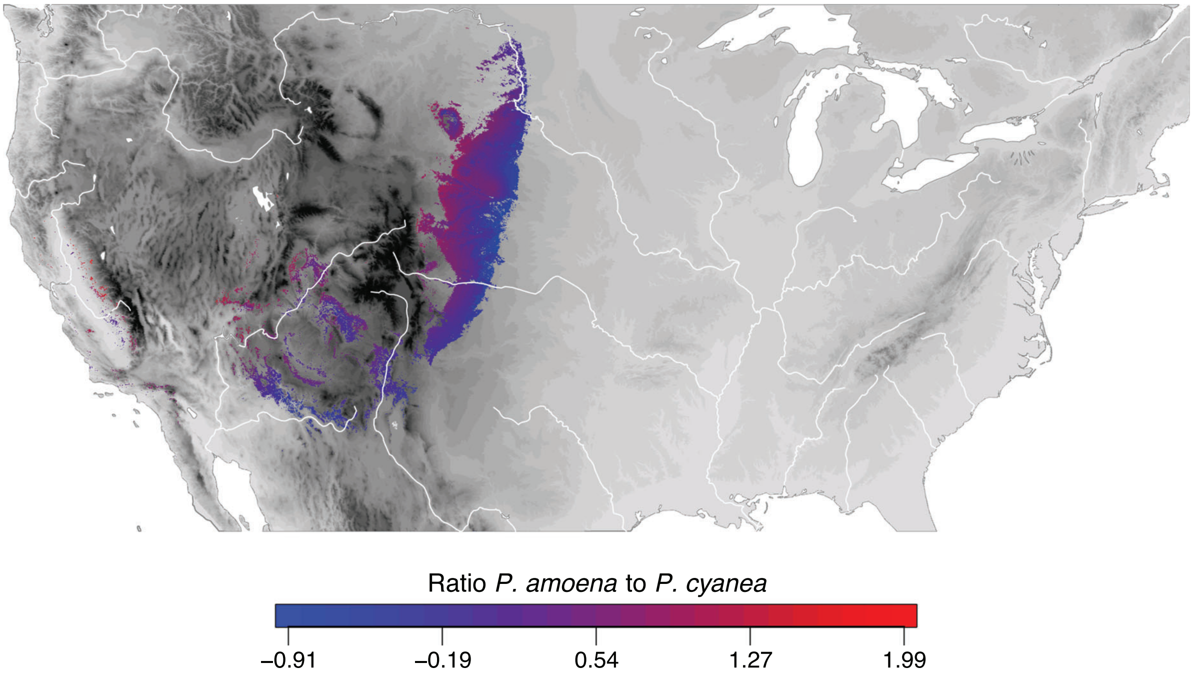

Figure 4. Log-transformed ratio of predicted parental species relative abundance within the inferred hybrid zone between *Passerina cyanea* and *P. amoena*. Within the area of overlap between *Passerina cyanea* and *P. amoena*, where both species have a minimum relative abundance of 0.05, this map depicts the gradual shift from *P. amoena* in the west to *P. cyanea* in the east.

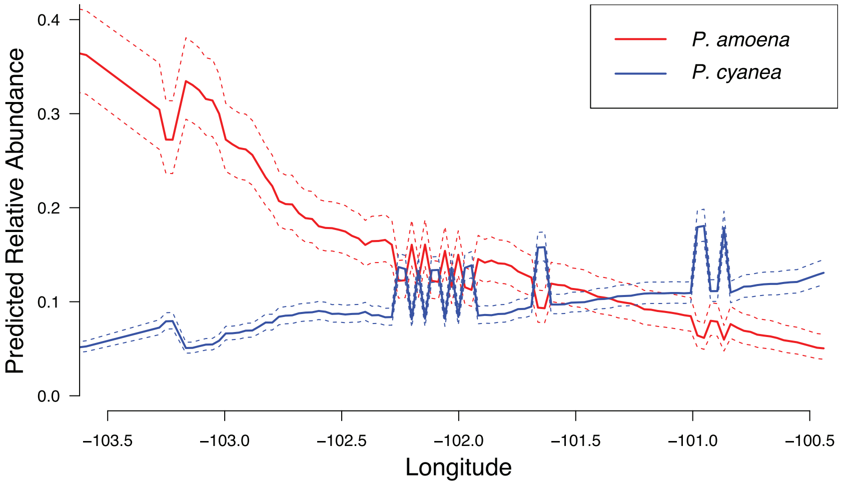

Figure 5. Predicted relative abundance of *Passerina cyanea* and *P. amoena* at 42.78125° latitude, approximately the mean latitude of the Niobrara River in the hybrid zone. Predicted relative abundances are based on eBird observations across the region where both species have a minimum relative abundance of 0.05. Observations in the eBird dataset indicate a more radical decline in the abundance of *P. amoena* from west to east than in *P. cyanea* from east to west. Both species also show spikes/dips when heterospecifics are similarly abundant, a pattern that may warrant investigation in the field.

of conspecific and heterospecific mates has no influence on the frequency of cross-species reproduction; (2) hybridization rates are highest where both species show roughly equal relative abundances, implying that prezygotic mate choice does little to prevent cross-species pairings; and (3) hybridization rates are highest where one parental taxon is abundant and the other is uncommon, implying that hybridization occurs when a representative of one species cannot find conspecific mates and must "settle" for heterospecifics. Moreover, we acknowledge that there are other avenues of hybridization research where eBird data alone are appropriate, but for our focus in this comment—estimating hybridization rates using only eBird data—we argue that estimated rates may be simplistic without integrating additional forms of data.

We believe these three hypotheses, and the above overall approach, will help future researchers bridge the gap from where eBird data are well suited to where we have demonstrated they are not. Ultimately, we emphasize that it is field studies of mate preference, encounter rates, pairing, extra-pair copulation, hybrid viability and vigor, and dispersal that are needed to fully understand hybridization rates for any species pair. Although eBird data are insufficient for estimating these parameters, they can be used as an important first step to help clarify how further studies of hybridization might proceed. In short, predicting (1) areas of overlap, where there is an opportunity for two species to hybridize, and (2) how abundance of parental taxa relates to hybridization are promising first steps in bridging this gap.

Conclusions

Here, we offer a suite of approaches to describe and predict the presence of hybrids and discuss the pros and cons of each approach. We argue that eBird data may best be used to make predictions about the possibility of hybridization, as these predictions allow the measurement of uncertainty, and empower citizen scientists to participate in avian hybrid research by groundtruthing predictions. Furthermore, we suggest that the data currently available from eBird are best suited to predicting areas of overlap between two hybridizing taxa, and thus can be used to predict opportunity to hybridize in a geographic context. Currently, available eBird data can be used to predict relative abundance and habitat suitability for hybrids specifically, but these predictions suffer from the limited number of hybrid observations. In these cases, museum and molecular data may be better suited for describing the frequency of hybrids, which we also demonstrate above. Finally, we offer an approach to predicting parental taxon abundance within predicted areas of overlap, thereby opening the door to future study without overextending the eBird dataset.

AUTHOR CONTRIBUTIONS

NRM synthesized results and wrote the manuscript, and also worked together with PJD to filter the eBird dataset. PJD wrote R scripts and performed analyses, generated figures, and contributed to writing the manuscript. MDC initiated the project, provided helpful insights and direction throughout, contributed to writing the manuscript, and oversaw its completion. SAT helped initiate the project, consulted on analyses, and contributed to writing the manuscript.

ACKNOWLEDGMENTS

We are indebted to all the birders who submit observations to eBird and all the collectors that deposit specimens in natural history museums, #collectionsareessential. All computing was done with the support of the Teton cluster (https://doi.org/10.15786/M2FY47), which is part of the University of Wyoming’s Advanced Research Computing Center. Data storage was provided by the University of Wyoming’s petaLibrary (https://doi.org/10.15786/G22X-KE07). We declare no conflict of interest.

DATA ARCHIVING

There are no data to be archived. The full eBird dataset is available upon request at https://ebird.org/data/download. All scripts used in this study are available and free to download at https://github.com/mcarling/avianhybridization.

LITERATURE CITED

Amatulli, G., S. Domisch, M.-N. Tuanmu, B. Parmentier, A. Ranipeta, J. Malczyk, and W. Jetz. 2018. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci. Data 5:180040.

Baiz, M. D., A. W. Wood, A. Brelsford, I. J. Lovette, and D. P. L. Toews. 2020. Pigmentation genes show evidence of repeated divergence and multiple bouts of introgression in Setophaga Warblers. Curr. Biol. 31:643–649.

Baker, M. C. 1991. Response of male Indigo and lazuli buntings and their hybrids to song playback in allopatric and sympatric populations. Behaviour 119:225–242.

Billerman, S. M., C. Cicero, R. C. K. Bowie, and M. D. Carling. 2019. Phenotypic and genetic introgression across a moving woodpecker hybrid zone. Mol. Ecol. 28:1692–1708.

Campagna, L., M. Repenning, L. F. Silveira, C. S. Fontana, P. L. Tubaro, and I. J. Lovette. 2017. Repeated divergent selection on pigmentation genes in a rapid finch radiation. Sci. Adv. 3:e1602404.

Carling, M. D., and H. A. Thomassen. 2012. The role of environmental heterogeneity in maintaining reproductive isolation between hybridizing Passerina (Aves: Cardinalidae) buntings. Int. J. Ecol. 2012. https://doi.org/10.1155/2012/295463

Carling, M. D., and B. Zuckerberg. 2011. Spatio-temporal changes in the genetic structure of the Passerina bunting hybrid zone. Mol. Ecol. 20:1166–1175.

Coyne, J. A., E. H. Kay, and S. Pruett-Jones. 2007. The genetic basis of sexual dimorphism in birds. Evolution. https://doi.org/10.1111/j.1558-5646.2007.00254.x.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24:332–340.

Drovetski, S. V., A. B. Reeves, Y. A. Red’kin, I. V. Fadeev, E. A. Koblik, V. N. Sotnikov, and G. Voelker. 2018. Multi-locus reassessment of a striking discord between mtDNA gene trees and taxonomy across two congeneric species complexes. Mol. Phylogenet. Evol. 120:43–52.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging?. John Wiley & Sons, Ltd.

Emlen, S. T., J. D. Rising, and W. L. Thompson. 1975. A behavioral and morphological study of sympatry in the Indigo and lazuli buntings of the Great Plains. Wilson Bull 87:145–302.

Everson, K. M., J. F. McLaughlin, I. A. Cato, M. M. Evans, A. R. Gastaldi, K. K. Mills, K. G. Shink, S. M. Wilbur, and K. Winker. 2019. Speciation, gene flow, and seasonal migration in *Catharus thrushes* (Aves:Turdidae). Mol. Phylogenet. Evol. 139:106564.

Friedl, M., and D. Sulla-Menashe. 2015. MCD12Q1 MODIS/Terra+ aqua land cover type yearly L3 global 500m SIN grid V006 [data set]. NASA EOSDIS L. Process. DAAC.

Funk, V. A. 1985. Phylogenetic patterns and hybridization. Ann. Missouri Bot. Gard 72:681.

Hewitt, G. M. 2008. Speciation, hybrid zones and phylogeography—or seeing genes in space and time. Mol. Ecol. 10:537–549.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. A J. R. Meteorol. Soc. 25:1965–1978.

Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. T. Kelling, and D. Fink. 2020. Analytical guidelines to increase the value of citizen science data: using eBird data to estimate species occurrence. bioRxiv. 574392.

Justen, H., A. A. Kimmitt, and K. E. Delmore. 2020. Estimating hybridization rates in the wild: easier said than done? Evolution 75:2137–2144.

Justyn, N. M., C. T. Callaghan, and G. E. Hill. 2020. Birds rarely hybridize: a citizen science approach to estimating rates of hybridization in the wild. Evolution 74:1216–1223.

Keele, L. 2007. Generalized additive models. Pp. 137–159 *in* Semiparametric regression for the social sciences. John Wiley & Sons, Ltd., Hoboken, NJ.

Linck, E., K. Epperly, P. Van Els, G. M. Spellman, R. W. Bryson, J. E. McCormack, R. Canales-Del-Castillo, and J. Klicka. 2019. Dense geographic and genomic sampling reveals paraphyly and a cryptic lineage in a classic sibling species complex. Syst. Biol 68:956–966.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol Evol 20:229–237.

Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? BioEssays 38:140–149.

Manthey, J. D., and M. B. Robbins. 2016. Genomic insights into hybridization in a localized region of sympatry between pewee sister species (*Contopus sordidulus* × *C. virens*) and their chromosomal patterns of differentiation. Avian Res. 7:6.

Mayr, E. 1963. Populations, species and evolution: an abridgement of animal and evolution. Harvard Univ. Press, Cambridge, MA.

Ottenburghs, J., and D. L. Slager. 2020. How common is avian hybridization on an individual level?. Evolution 74:1228–1229.

Ottenburghs, J., R. H. S. Kraus, P. van Hooft, S. E. van Wieren, R. C. Ydenberg, and H. H. T. Prins. 2017. Avian introgression in the genomic era. Avain Res. 8:1–11.

Pyle, P., W. A. Leitner, L. Lozano-Angulo, F. Avilez-Teran, H. Swanson, E. G. Limón, and M. K. Chambers. 2009. Temporal, spatial, and annual variation in the occurrence of molt-migrant passerines in the Mexican monsoon region. Condor 111:583–590.

Rohwer, S., L. K. Butler, D. R. Froehlich, R. Greenberg, and P. P. Marra. 2005. Ecology and demography of east–west differences in molt scheduling of Neotropical migrant passerines. Pp. 87–105 in R. Greenb and P. P. Marra, eds. Birds of two worlds: the ecology and evolution of migration. Johns Hopkins Univ. Press, Baltimore, MD.

Seehausen, O. 2004. Hybridization and adaptive radiation. Trends Ecol Evol. 19:198–207.

Strimas-Mackey, M., E. Miller, and W. Hochachka. 2018. auk: eBird data extraction and processing with AWK.

Strimas-Mackey, M., W. M. Hochachka, V. Ruiz-Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. Kelling, D. Fink, and A. Johnston. 2020. Best practices for using eBird data. Version 1.0. Cornell Lab of Ornithology, Ithaca, NY.

Sullivan, B. L., J. L. Aycrigg, J. H. Barry, R. E. Bonney, N. Bruns, C. B. Cooper, T. Damoulas, A. A. Dhondt, T. Dietterich, A. Farnsworth, et al. 2014. The eBird enterprise: an integrated approach to development and application of citizen science. Biol. Cons. 169:31–40.

Swenson, N. G. 2006. Gis-based niche models reveal unifying climatic mechanisms that maintain the location of avian hybrid zones in a North American suture zone. J. Evol. Biol. 19:717–725.

Taylor, S. A., and E. L. Larson. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nature Ecol Evol. 3:170–177.

Taylor, S. A., T. A. White, W. M. Hochachka, V. Ferretti, R. L. Curry, and I. Lovette. 2014. Climate-mediated movement of an avian hybrid zone. Curr. Biol. https://doi.org/10.1016/j.cub.2014.01.069.

Thompson, K. A., M. Urquhart-Cronish, K. D. Whitney, L. H. Rieseberg, and D. Schluter. 2020. Patterns, predictors, and consequences of dominance in hybrids. Am. Nat. 197. https://doi.org/10.1086/712603.

Vallender, R., S. L. Van Wilgenburg, L. P. Bulluck, A. Roth, and R. Canterbury. 2009. Cryptic hybridization in the golden-winged warbler (*Vermivora chrysoptera*). Avian Conserv. Ecol 4:art4.

Willis, P. M., M. J. Ryan, and G. G. Rosenthal. 2011. Encounter rates with conspecific males influence female mate choice in a naturally hybridizing fish. Behav. Ecol 22:1234–1240.

Wood, S., and M. S. Wood. 2015. Package 'mgcv.' R Packag. version 1:29.

Young, B. E. 1991. Annual molts and interruption of the fall migration for molting in lazuli buntings. Condor 93:236–250.

Associate Editor: R. C. Fuller
Handling Editor: T. Chapman

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Hybrid proportions for 11 species pairs that are likely represented in the eBird dataset. We calculated proportions for each year separately as well as for all years combined. Additionally, we calculated proportions using the data from across the range as in Justyn et al. (2020). We also calculated hybrid proportions using two methods that restricted the datasets such that only data from the hybrid zone were used.

Figure S1. Predicted area of breeding range overlap for *Passerina cyanea* and *P. amoena*, where both parental species have a predicted relative abundance of at least 0.1. Shading in underlying maps represent elevation, and white lines trace prominent rivers.

Figure S2. Predicted area of breeding range overlap for *Passerina cyanea* and *P. amoena* where parental species have a predicted relative abundance of at least 0.05. Shading in underlying maps represent elevation, and white lines trace prominent rivers.

Figure S3. Predicted area of breeding range overlap for *Passerina cyanea* and *P. amoena* where parental species have a predicted relative abundance of at least 0.025. Shading in underlying maps represent elevation, and white lines trace prominent rivers

Figure S4. Predicted area of breeding range overlap for *Passerina cyanea* and *P. amoena* where parental species have a predicted relative abundance of at least 0.01. Shading in underlying maps represent elevation, and white lines trace prominent rivers.

Figure S5. Comparison of hybrid proportions in the *Sphyrapicus* hybrid zone based on eBird and molecular data.

Figure S6. Predicted relative abundance of *Passerina cyanea* and *P. amoena* at 41.01042° latitude, approximately the mean latitude of the Platte River in the hybrid zone