Can genomic data alone tell us whether speciation happened with gene flow?

Ming Yang1 | Ziwen He1 | Suhua Shi1 | Chung-I Wu1,2

1State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China
2Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA

Correspondence
Suhua Shi and Chung-I Wu, Life Science, Sun Yat-Sen University, Guangzhou, China Ecology and Evolution, University of Chicago, Chicago USA. Emails: lssssh@mail.sysu.edu.cn and ciwu@uchicago.edu

Funding information
National Natural Science Foundation of China, Grant/Award Number: 91331202, 41103208, 31600182; The 985 Project, Grant/Award Number: 3300-31131105; National Key Research and Development Plan, Grant/Award Number: 2017FY100705

Abstract
The allopatric model, which requires a period of geographical isolation for speciation to complete, has been the standard model in the modern era. Recently, "speciation with gene flow" has been widely discussed in relation to the model of "strict allopatry" and the level of DNA divergence across genomic regions. We wish to caution that genomic data by themselves may only permit the rejection of the simplest form of allopatry. Even a slightly more complex and realistic model that starts with subdivided populations would be impossible to reject by the genomic data alone. To resolve this central issue of speciation, other forms of observations such as the sequencing of reproductive isolation genes or the identification of geographical barrier(s) will be necessary.

KEYWORDS
genomics/proteomics, hybridization, molecular evolution, population genetics, theoretical

In the modern synthesis, speciation facilitated by allopatry (often referred to simply as allopatric speciation) has been the standard model for more than half a century (Coyne & Orr, 2004; Mayr, 1963). In this view, speciation requires a period of geographical isolation during which the diverging populations complete the process of speciation without experiencing gene flow. Although this period of strict allopatry is essential, events before and after this period may vary from species to species, potentially leading to false rejections of the allopatric model.

The conceptual basis for requiring a period of strict isolation is stated in Mayr (1963). Gene flow is perceived as a homogenizing force, and even a tiny amount is assumed to be able to reverse the divergence. In this view, all genic regions gradually evolve functional characteristics that are cohesive within the same species but each region would have negative interactions with the genome of the diverging population. While epistasis in fitness is commonly accepted in the modern synthesis (Dobzhansky, 1937; Wright, 1940), strict allopatry postulates genomewide "cohesiveness" (Wu, 2001).

In the modern view of genetics, many large segments of the genome may harbour no loci of fitness consequence between nascent species (Fontaine et al., 2015; Malinsky et al., 2015; Poelstra et al., 2014; Wu & Ting, 2004; Zhang, Dasmahapatra, Mallet, Moreira, & Kronforst, 2016). These neutral gene regions can easily travel between nascent species during speciation, while selection operates only on genes that have diverged functionally. In this genic view (Feder, Egan, & Nosil, 2012; Seehausen et al., 2014), gene flow would not retard ecological divergence when different alleles are selected against in different environments. The implication is that gene flow would pose little problem for species divergence and strict allopatry is not necessarily the dominant mode of speciation (Feder, Flaxman, Egan, Comeault, & Nosil, 2013; Schluter, 2000; Schluter & Conte, 2009).

Despite increasing discussions of this alternative model in the era of genomics, strict allopatry may still be necessary if reproductive isolation (RI) is taken into account. Consider the simple Dobzhansky–Muller model of hybrid incompatibility where the initial state is a two locus (a, b) haplotype. Let a and b evolve towards A and B, respectively, under positive selection but A and B are incompatible. In full geographical isolation, population 1 may evolve to (A, b), while population 2 evolves to (a, B) and hybrid incompatibility would ensue. If gene flow occurs between the two diverging populations, the mutual interference between a → A and b → B would render each pathway impassable, effectively stopping the evolution of postmating isolation. These dynamics, first explored in the context of gene silencing after
gene duplication, are highly stochastic (Li, 1980; Maruyama & Takahata, 1981; Watterson, 1983). Although the same dynamics have been analysed for the evolution of hybrid incompatibility only in deterministic terms (Agrawal, Feder, & Nosil, 2011; Bank, Bürger, & Hermisson, 2012; Nosil & Flaxman, 2011), it is clear that gene flow can strongly retard the evolution of reproductive incompatibility.

Therefore, depending on the nature of genic interactions, gene flow can be seen either as a neutral force or as a strong retardant of speciation. "Strict allopatry" versus "speciation with gene flow" obviously must be resolved by empirical observations (Faria et al., 2014; Seehausen et al., 2014; Sousa & Hey, 2013). It has been optimistically suggested that the torrent of genomic data will provide a clear answer. In this short note, we raise caveats against this optimism. It is unlikely that the issue of "speciation with gene flow" will be resolved by genomic data alone, regardless of the sophistication of the statistical tools (Becquet & Przeworski, 2009; Sousa, Grelaud, & Hey, 2011; Strasburg & Rieseberg, 2010, 2011, 2013) or the amount of genomic data. Other types of observations will be needed in conjunction with the genomic data.

In previous tests of strict allopatry, the model is the Level 1 model of allopatry shown in Figure 1a. An uninterrupted period of geographical isolation, during which gene flow ceases, is portrayed between T1 and T2. Speciation is completed by T2 after which the species may come into contact again. Tests of allopatry are usually based on this Level 1 model because even a slightly more complex model will be nearly impossible to reject, if it is testable at all, by the genomic data alone (see Figure 1b and text below) (Ellegren et al., 2012; Garrigan et al., 2012; Geraldès, Basset, Smith, & Nachman, 2011; Geraldès, Ferrand, & Nachman, 2006; Herrig, Modrick, Brud, & Llopart, 2014; Kronforst et al., 2013; Leaché, Harris, Maliska, & Linkem, 2013; Malund et al., 2012; Muñoz et al., 2013; Nadachowska & Babik, 2009; Niemiller, Fitzpatrick, & Miller, 2008; Osborne, Batstone, Hiscock, & Filatov, 2013; Runemark, Hey, Hansson, & Svensson, 2012; Sousa, Carneiro, Ferrand, & Hey, 2013; Winker, Mccracken, Gibson, & Peters, 2013; Won, Sivasundar, Wang, & Hey, 2005).

In statistical terms, the Level 1 model (Figure 1a) is an excellent null hypothesis that assumes that all genomic regions have diverged more quickly than others. The rationale for testing such a simple model is that the failure to reject it may also mean that strict allopatry in general cannot be rejected (Innan & Watanabe, 2006; Presgraves & Yi, 2009; Yamamichi, Gojobori, & Innan, 2012). This is because complex models with more parameters would be even more difficult to reject. Importantly, given massive genomic data and refined statistical methods, there should be sufficient statistical power to reject the null model if it is indeed incorrect.

The issue arises because the null model of Figure 1a is often rejected (Garrigan et al., 2012; Kronforst et al., 2013; Leaché et al., 2013; Muñoz et al., 2013; Osaka & Wu, 2005; Osborne et al., 2013; Pinho & Hey, 2010). Highly divergent regions, in the form of "genomic islands" that are not expected by the null model, have been widely reported (Carneiro et al., 2014; Ellegren et al., 2012; Malinsky et al., 2015; Poelstra et al., 2014; Toews et al., 2016; but see Burri et al., 2015; Pennisi, 2014; Renaut et al., 2013). Although some results may have been misinterpreted due to low polymorphism instead of high divergence (Cruickshank & Hahn, 2014; Payseur & Rieseberg, 2016; Wolf & Ellegren, 2017), many are likely correct in rejecting the null model of Figure 1a.

The rejection of Level 1 allopatry thus raises the bar—"Can a more realistic model of allopatric speciation be rejected?" Figure 1b shows a Level 2 model of allopatric speciation. Unlike in the Level 1 model whereby geographical isolation is imposed on a geographically panmictic population, full isolation in nature may often happen between partially differentiated geographical populations. Indeed, many studies following the Takahata, Satta, and Klein (1995) analysis have found the ancestral population to have an effective size much larger than the extant one (Mailund, Munch, & Schierup, 2014; Osaka & Wu, 2005; Satta, Hickerson, Watanabe, O'Higgins, & Klein, 2004; Won & Hey, 2005; Zhou et al., 2007). A plausible explanation for the very large effective size would be populations with deep structure (Becquet & Przeworski, 2009; Osaka & Wu, 2005; Zhou et al., 2007), hence supporting the model of Figure 1b.

If allopatry is indeed the most common mode of speciation and generally follows the Level 2 model, genomic data could not distinguish between "strict allopatry" and "speciation with gene flow". Note that genomic data by themselves can only inform whether there has been gene flow between T0 (when the species started to differentiate) and T2 (when speciation was completed). They cannot tell us whether there exists a period (T1–T2) without gene flow. Unless the pattern of geographical differentiation at T1 is known, it will not be possible to test the Level 2 model of Figure 1b. Of course, methods exist for inferring the ancient population structure but such models (Innan & Watanabe, 2006; Takahata & Satta, 1997; Takahata et al., 1995; Yang, 2002) have to assume a period of strict allopatry, akin to the Level 2 model. In other words, either allopatry is true or the ancient population structure has to be known to test the Level 2 model. Genomic data alone are therefore insufficient to reject such a model.

Not knowing the ancient population structure raises other conceptual issues. It has been suggested that population differentiation between T0 and T1 in Figure 1b should be considered "incipient speciation", implying that the Level 2 model is also a model of "speciation with gene flow" (Butlin, 2010). As stated earlier, the models of allopatric speciation require a period of strict isolation with no gene flow (T1–T2 in Figure 1). Thus, while the rejection of allopatry entails the proof that such a period is absent, other events before T1, or after T2, are peripheral to the allopatric debate. Furthermore, it has been demonstrated that when the differentiation at T1 has little to do with incipient speciation (e.g., in simulations of neutral divergence), allopatric speciation simulated by the Level 2 model could still be interpreted as "speciation with gene flow" (Pinho & Hey, 2010). To take the simulation approach a step further, we suggest a prospective approach starting with genomic data of actual geographical populations (such as the extant human populations). We will then be able to test whether the simulated new species, when analysed retrospectively, would falsely reject the model of allopatry. The approach is based on realistic
extant populations and should be a useful tool for analysing various allopatric models of speciation.

More complex models of allopatry (beyond Level 2) have been applied to many taxa (Christe et al., 2017; Duvaux, Belkhir, Boulesteix, & Boursot, 2011; Filatov, Osborne, & Papadopulos, 2016; Le Gac et al., 2016; Lohse, Clarke, Ritchie, & Etges, 2015; Nada-
chowska-Brzyska et al., 2013; Roux, Tsagkogeorga, Bierne, & Galtier, 2013; Tine et al., 2014). These are usually models with extensive recent gene flow applied to “species pairs” that have not developed full RI. For these taxa, the issue is not gene flow. Instead, it is whether the taxa will need a long period of strict allopatry in the future to develop full RI. (One may also ask, retrospectively, whether the absence of full RI might be connected to insufficiently long period of geographical isolation in the past.)

It seems clear that information in addition to genomic data will be necessary for testing the more realistic models of allopatric speciation. There are at least two options. First, we may use functional genomic means to test the general allopatric models. In allopatry, the divergence is a function of the timing of erecting the geographical isolation, and “speciation genes” underlying RI should be no more divergent than the rest of the genome. In speciation with gene flow, true speciation genes that evolve in the very incipient stage should be older than the others (Ting, Takahashi, & Wu, 2001).

Second, we suggest that the biogeographical records of species in question be incorporated into the analysis of genomic data. It would seem desirable to analyse species pairs that are separated by a well-defined geographical barrier as the geological record of the barrier can guide tests of gene flow with timing information. The Isthmus of Panama is the best-known example. Its gradual and irreversible closure has made speciation between the Pacific and Atlantic sides the best examples of allopatry (Bacon et al., 2015; Palumbi, 1994). Besides permanent barriers, nonpermanent barriers exist in many terrestrial, aquatic and marine environments. Speciation associated with such barriers may be informative about the roles of gene flow. The Indo-Pacific Barrier near the Strait of Malacca could be such an example. In addition, a nonpermanent barrier may continue to produce new species, hence providing valuable data on speciation events of different ages.

Whether speciation can proceed in the presence of gene flow is important in many aspects of biology including genetics, development, behaviour, ecology and biogeography. On one hand, it may seem difficult to envisage the emergence of thousands of species of beetles in the Amazon, all requiring periods of strict allopatry (Erwin, 1996; Hoorn et al., 2010). On the other hand, the model of “speciation with gene flow” will require the convincing rejection of strict allopatry, which remains challenging even with the massive genomic data.

ACKNOWLEDGEMENTS

We thank Trevor Price, Marcus Kronforst, Roger Butlin, Patrik Nosil and Loren Rieseberg for suggestions and discussions. This work was supported by the National Natural Science Foundation of China (grant numbers 91331202, 41130208 and 31600182); The 985 Project (grant 33000-31131105); National Key Research and Development Plan (2017FY100705).

AUTHOR CONTRIBUTIONS

C-I.W. wrote the manuscript with input from M.Y. All authors edited the manuscript.

DATA ACCESSIBILITY

No data needed in the main text.
REFERENCES


---