Divergence with Gene Flow: Models and Data

Catarina Pinho[1] and Jody Hey[2]

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto. Campus Agrário de Vairão, 4485-661 Vairão, Portugal; email: catarina@mail.icav.up.pt

[2]Department of Genetics, Rutgers University, Piscataway, New Jersey 08854; email: hey@biology.rutgers.edu

Annu. Rev. Ecol. Evol. Syst. 2010. 41:215–30

First published online as a Review in Advance on August 19, 2010

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

This article's doi:
10.1146/annurev-ecolsys-102209-144644

1543-592X/10/1201-0215$20.00

Key Words

disruptive selection, divergence population genetics, isolation with migration, linkage disequilibrium, speciation, species, sympatry

Abstract

Since Darwin first proposed that new species could arise without geographic separation, biologists have debated whether or not divergence occurs in the presence of gene exchange. Today we understand that new species can diverge while exchanging genes, depending on the strength of disruptive natural selection and the factors that affect the linkage relationships of genes under disruptive selection. This mode of diversification—divergence with gene flow—includes sympatric speciation, in which gene exchange occurs since onset of divergence, and secondary contact following a period of geographic isolation, as well as all sorts of situations in which gene flow happens intermittently. In recent years, statistical tools have been developed that can reveal the action of gene flow during divergence. Isolation-with-migration (IM) models include parameters for population size, time of population separation, and gene exchange, and they have been used extensively to estimate levels of gene exchange. A survey of studies that have used these models shows that a plurality find little evidence of gene flow; however, many report nonzero gene exchange.

INTRODUCTION: THE DARWIN/WAGNER DEBATE ON SPECIATION WITH GENE FLOW

Gene flow: the movement of alleles between partially separated populations caused when individuals who have one parent from each population reproduce and pass genes from one population into the other

Linkage disequilibrium: nonrandom association of alleles among different genes, such that some alleles tend to be found linked to particular alleles at other genes

As clear and persuasive as Darwin was on natural selection and adaptation in *On the Origin of Species . . .* , he was uncharacteristically difficult to understand when, in Chapter IV, he explained how one species gives rise to two. Over several pages, Darwin walked the reader through his only figure and described how natural selection could create multiple species from a single species that inhabits a continuous large area, without the geographic separation of populations. Darwin called his model the principle of divergence, but exactly how it is supposed to work has sometimes eluded readers (Gould 2002, Tammone 1995).

Darwin's emphasis on his principle of divergence (what today we would identify as a model of sympatric speciation) invoked a strong critique by Moritz Wagner (1873), who argued that migration and reproduction across the range of the species (what we would today identify as gene flow) would impede the divergence process. Darwin acknowledged this point in the sixth edition of this book but maintained that his model was appropriate for most cases of speciation.

It is striking that the Darwin/Wagner debate anticipated, by roughly 100 years, the energetic debate over sympatric speciation that emerged in the late twentieth century. Neither had an inkling of genes or even the rudiments of an appropriate inheritance model, and yet Wagner put his finger precisely on that factor that can prevent divergence. If Darwin's model was to be correct, and if his principle of divergence truly described a mechanism for the origin of species, then it must be possible (and common) for two species to arise from one species in the presence of gene exchange.

In this review of gene exchange and divergence, we begin by developing a general genetic perspective on how populations become genetically differentiated in the presence of gene flow. The literature on this process is quite extensive [see recent reviews by Bolnick & Fitzpatrick (2007), Butlin (2005), Coyne & Orr (2004), Gavrilets (2003), and Via (2001)]; however, by focusing narrowly on the factors shaping linkage disequilibrium, we hope to present a simplified and accessible summary. We then turn to a review of recent methodological advances and applications to see what has been learned in recent years about the frequency and magnitude of gene exchange during divergence.

SPECIATION IS A MULTIGENE PROCESS

We conclude that speciation has occurred if two closely related populations are genetically incompatible in a way that prevents hybrids from being formed or that causes low hybrid fitness. But just how much divergence is required to create a genetic barrier? At minimum, we can say that a reproductive barrier requires that populations be different at two genes (Dobzhansky 1937). It is difficult, though not impossible (Orr 1991), to explain speciation and low hybrid fitness with a model in which two populations are fixed for different alleles at just a single gene. This is because a difference between populations at one gene, for alleles that cause low fitness in a heterozygote, is unlikely to come to pass simply because any new allele that is harmful in heterozygous condition is quickly lost soon after it arises.

Bateson (1909), Dobzhansky (1937) and Muller (1940) independently suggested that more than one gene is required for a new species and that the genetic basis of reproductive isolation is likely to involve at least two interacting loci (Orr 1996). The Bateson-Dobzhansky-Muller (BDM) model begins with recently separated populations that are both completely homozygous for allele *A* at one locus and for allele *B* at another locus (all individuals have genotype *AABB*). Then at the first locus, one population has a mutation to allele *a* that replaces all other gene copies at this locus,

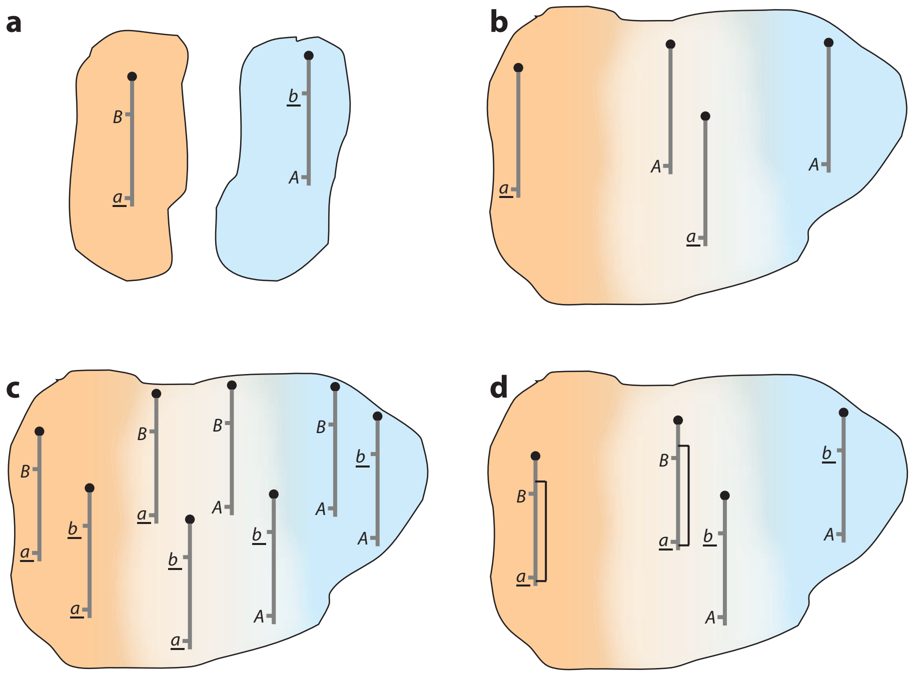

Figure 1

A genetic perspective on divergence. Two loci, A and B, each have two alleles favored in different populations or portions of the range. For simplicity, both genes are represented as occurring on a single chromosome. Chromosomal genotypes are shown in the population or geographic region where they are common. (*a*) Two allopatric populations, each fixed for different alleles at the two loci. (*b*) Divergence under disruptive selection that leads to geographic differentiation at both genes. The range of the species includes areas at the extremes of the selective gradients, in which only a single chromosome is common (on the *left* and *right*), as well as a central region where all four possible chromosomes may occur. The numbers of each chromosome type in the central region depend on the strength of disruptive selection, levels of assortative mating, and rates of recombination in individuals that are heterozygous at both genes. (*c*) If heterozygotes are not rare and recombination breaks down linkage disequilibrium between the A and B genes, then polymorphisms at other genes not under selection may be widespread throughout the range of the species. (*d*) If recombination is low or is suppressed in heterozygotes (indicated by a black bar representing a chromosomal rearrangement on one chromosome type), then other genes will also come into linkage disequilibrium with the selected genes.

whereas in the second population at the second locus, a new allele *b* replaces all other copies of the gene. **Figure 1*a*** shows the two chromosome types that would be found in their respective populations (assuming, for simplicity, that both loci are on the same chromosome). If a hybrid is produced, it will be heterozygous at both loci and will have genotype *AaBb*. But alleles *a* and *b* have never been together on the same individual before, hence their interaction has never been subject to the scrutiny of selection. It is possible that they interact negatively, causing the hybrids to have reduced or possibly zero fitness, in which case speciation would be complete. If two genes are not sufficient and hybrid fitness is only partly reduced, we can then imagine adding more loci to the model with new alleles becoming fixed in just one population and further contributing to low hybrid fitness.

Divergence Without Geographic Separation

Disruptive selection (diversifying selection): when selection favors individuals with traits at the opposite ends of a distribution and removes individuals with intermediate traits

The BDM model invokes two mechanisms for preventing the spread of alleles that have become fixed in one population into the other population. The first is geographic separation, and the second, which kicks in when hybrids are produced, is epistatic incompatibility between alleles that have become fixed in different populations.

The flip side of the BDM model is that if hybrids are produced and are not completely sterile, then it may happen that an allele that is fixed in one population does spread through the other population. When this occurs, then both the actual amount of divergence and the potential for epistatic incompatibilities are reduced.

These ideas can be used when we turn to the case of two populations that are not isolated from each other, but that overlap or abut each other and that exchange genes with each other. Although the BDM model does not apply directly to this case, we learn from the BDM model that if the two populations are to diverge in the face of gene exchange, then (*a*) there must be forces acting against the spread of genes across both populations, and (*b*) there must be forces that lead to a stronger barrier to gene exchange and that allow divergence to increase.

Natural selection can certainly create some divergence, at least at the level of individual genes, within an interbreeding population. For example, **Figure 1*b*** depicts a single population in which a new allele *a* is favored and replaces the *A* allele, but only over a portion of the population's range (that is, the *A* allele is favored over the remainder of the range). In this kind of model, the disruptive selection itself is the direct cause of the limited spread of alleles, and there are no physical barriers to gene movement, nor is there an epistatic cause of low hybrid fitness as in the BDM model. Rather, low fitness individuals occur when specific genotypes occur in the wrong part of the range of the population.

An investigator coming upon the situation in **Figure 1*b*** would perceive a single species that has a strong reciprocal shift in *A* and *a* allele frequencies across the range. Other polymorphic loci that are not under disruptive selection cannot be expected to show the same regional limitations. For example, **Figure 1*c*** shows the entire range of the species being polymorphic for two alleles, *B* and *b*, at a second locus that, in this case, is not under disruptive selection.

However, the situation in **Figure 1*b*** can be the seeds for continuing divergence that spreads across the genome to affect more than just the genes that are initially under disruptive selection. To see how this works, suppose that the B locus is actually closely linked to the A locus and that the *b* allele arises on a chromosome bearing the *A* allele. Then, for two reasons, we may not expect to find *b* alleles on the same chromosomes as *a* alleles. First, the opportunity for recombination, which could move a *b* allele onto a chromosome carrying an *a* allele, is low because the geographic distribution of the *A* and *a* alleles, due to disruptive selection, means that there are fewer *Aa* heterozygotes than would be expected under random mating. Second, even when *Aa* heterozygotes occur, the tight linkage between the two loci means that recombination events are rare. In short, the disruptive selection on the *A* and *a* alleles can cause divergence at linked loci, and it can cause a shortage of some gene combinations (that is, *ab* and *AB* chromosomes) relative to that expected under random mating and free recombination. This shortage is called linkage disequilibrium, and it is a key prerequisite for speciation in most models of divergence in the absence of a geographic barrier (Bolnick & Fitzpatrick 2007, Felsenstein 1981, Gavrilets 2003, Rice & Hostert 1993).

The central point here is that disruptive selection on a locus essentially creates a place in the genome where additional divergence can accumulate. The length of this region depends on the recombination rate around the selected locus, but more importantly the selection regime itself can directly reduce the number of heterozygotes that are required for recombination to breakdown linkage disequilibrium.

In the absence of geographic separation, disruptive selection must be an ultimate cause of divergence. However, there are other key factors that come into play because of their effect on the level of linkage disequilibrium around the gene or genes that are the targets of disruptive selection. For example, initial linkage disequilibrium can occur when two geographically separated populations, which have diverged slightly due to selection or drift acting differently upon each population, come back into contact. Upon coming into contact and exchanging genes, the divergence process may possibly continue if disruptive selection continues to act sufficiently to maintain linkage disequilibrium.

Another important factor that can facilitate low recombination is assortative mating associated with the traits that are the subject of disruptive selection. In such circumstances, individuals tend to mate with others that share their genotype because of indirect effects of the alleles that are under disruptive selection. This could be simply due to the geographic structuring that occurs because of disruptive selection, so that most individuals come into contact only with others of the same genotype (Kirkpatrick & Ravigné 2002), or it could be because of a pleiotropic effect in which the alleles under disruptive selection also affect other traits related to mate choice. Probably the most discussed example of this, which applies to many phytophagus insects, is when genetic variation that affects food choice also leads to assortative mating because organisms tend to mate at feeding sites (Bush 1994).

Yet another major factor can be a chromosomal rearrangement that is itself in linkage disequilibrium with the locus under disruptive selection. This is represented in **Figure 1*d*** as a black bar that occurs on chromosomes with the *a* allele and that spans the A and B loci. Such a suppressor of recombination can have a major impact on the divergence process. In the first place, the low recombination and the regime of disruptive selection causes and sustains strong linkage disequilibrium around the A locus. Second, this disequilibrium comes to affect other loci that lie within the span of the rearrangement. Third, additional alleles at other loci that interact with the polymorphisms at the A locus can arise and contribute to an even stronger regime of disruptive selection. To understand this, suppose that allele *b*, which is linked to allele *A*, is favored over exactly the same range as allele *A*. This could be because this allele is suited to the same particular regime of disruptive selection (e.g., as a modifier of the effects of the *A* allele), or it could be that allele *b* reduces the rate at which individuals with allele *A* chromosomes reproduce with individuals bearing chromosomes with the *a* allele. If it is the latter, then we have the beginnings of a process of reinforcement of a reproductive barrier and an increase in assortative mating.

The stronger the reduction in recombination or the more genes affected, then the more opportunity there is for the recruitment of variation at additional loci that respond to the regime of disruptive selection; thus, the faster will the two populations diverge and the sooner will speciation be complete. The association can also act to limit the effects of selective sweeps, which would contribute to the similarity of the two populations by hindering the spread of alleles occuring in the region of reduced recombination that would be favored in both populations (Navarro & Barton 2003). In effect, the reduced recombination that is represented in **Figure 1*d*** plays a role that is analogous to that of the population separation in **Figure 1*a***. They both reduce the segregation of recombination chromosomes and the spread of alleles between the populations.

FORCES IN CONFLICT: DISRUPTIVE SELECTION VERSUS GENE FLOW AND RECOMBINATION

We can appreciate gene flow and diversifying natural selection as two forces in opposition, with each force tending to limit the other. On the one hand, we have disruptive selection and whatever factors contribute to linkage disequilibrium among loci that are differentiated by the selective regime. But on the other hand, gene exchange and recombination reduce the linkage disequilibrium that can, in turn, prevent the recruitment of additional loci to the divergence process.

Will gene exchange prevent divergence? Not necessarily, because strong selection may prevent some alleles from passing between populations whereas other unlinked genes that are not affected by selection can pass between populations. In this way, there may be genes that effectively do not experience gene exchange and behave as islands of differentiation in an otherwise uniform genome (Wu 2001).

Evidence of a Role for Reduced Recombination in Divergence with Gene Flow

Models that invoke a role for reduced recombination in divergence with gene flow predict that the genetic targets of disruptive selection or genes that cause low hybrid fitness are more likely to be found near chromosomal regions with reduced recombination (Navarro & Barton 2003, Noor et al. 2001b, Rieseberg 2001, Trickett & Butlin 1994). In the case of the divergence of *Drosophila pseudoobscura* and *D. persimilis*, the genes associated with partial reproductive barriers are indeed linked to chromosomal inversions that differentiate the species (Noor et al. 2001a). The idea is that among populations that have diverged to some extent in allopatry, and that then come together and exchange genes, those that differ from chromosomal inversion differences are less likely to merge back into a single species. Consistent with this is the observation in the genus *Drosophila* that young pairs of sympatric species are more likely to differ by chromosomal inversions than are species pairs that are allopatric (Noor et al. 2001b).

If divergence with gene flow is facilitated by reduced recombination, then we also expect to see evidence of more gene flow in parts of the genome where recombination is high. These types of patterns have been found to be associated with chromosomal breakpoints and inversions in the apple maggot *Rhagoletis pomonella* (Feder 2003), in *D. pseudoobscura* and *D. persimilis* (Machado et al. 2002), and in the sunflowers *Helianthus annuus* and *H. petiolaris* (Yatabe et al. 2007). Low gene flow relative to other parts of the genome has also been found in low recombination areas near chromosomal centromeres between different forms of the mosquito *Anopheles gambiae* (Slotman et al. 2006, Stump et al. 2005) and between subspecies of the european rabbit *Oryctolagus cuniculus* (Carneiro et al. 2009).

Gene Flow Mapping: Identifying Targets of Disruptive Selection by Studying Gene Flow

The intertwined and oppositional relationship between gene flow and natural selection means that we can study one of these mechanisms to learn more about the other. For an investigator trying to reconstruct how divergence occurred for a particular pair of sister species the first question is often whether or not gene flow occurred during the process. If it is found that gene flow has occurred, then the investigator is placed in the paradoxical but useful position of studying the forces driving divergence by identifying loci that have not diverged. Like an artist working with negative images, the investigator may not be able to identify or directly study those loci most targeted by selection, but by studying the gene exchange at other loci, the action of disruptive selection can be revealed. Evaluating which loci have experienced gene exchange and which have not can help to identify genomic locations that are contributing more than the average to divergence, as well as to reveal aspects of the genetic architecture that seem to be important for speciation (Geraldes et al. 2006, Machado et al. 2002, Rieseberg et al. 1999, Turner et al. 2005, Wu 2001).

DETECTING AND MEASURING GENE FLOW

Historically, before recent advances in population genetic methods, an inference that gene exchange was occurring during the divergence process usually depended on the biogeographic circumstances of the populations concerned. If sister species or members of a species complex are sympatric, and if they are entirely restricted to a small geographic area, then a conclusion of gene flow during species formation may seem parsimonious. The most famous such case is that of the cichlid fishes of the great African lakes (Tanganyika, Victoria, and Malawi), each of which harbors an endemic group of hundreds of closely related species (Kocher 2004, Salzburger & Meyer 2004).

If genetic data are available, such as allele or haplotype frequencies at some genes for each of two closely related species, then population genetic methods can be brought to bear on gene flow questions. However, estimates and conclusions regarding gene flow during divergence remain difficult to come by. The challenge is to understand what kinds of patterns of genetic variation are left in the genes of diverging species under different kinds of histories. For example, one of the main lessons to emerge from studying the population genetics of divergence is that a finding of shared variation between two species is not necessarily a signal of gene exchange. When a population splits into two, genetic variation continues to be shared by the daughter populations for a period of time thereafter (possibly a long period of time if populations are large and genetic drift is slow), even in the absence of gene exchange. As divergence proceeds, loci that were polymorphic in the ancestral population experience fixation of alleles in the descendant populations, and this sorting of gene lineages is part of the way the populations become different. The question of gene exchange often comes down to the challenging task of distinguishing among the two main causes of similarity between the genes from different populations: (*a*) incomplete lineage sorting since divergence began and (*b*) actual gene exchange.

Gene Flow Can Cause a High Variance in Divergence among Loci

Gene exchange during divergence can add significantly to the normal variance that occurs among genes in levels of shared polymorphism between diverging populations. Genes that are linked to those under disruptive selection may not experience any gene exchange, whereas other unlinked genes may experience substantial gene flow. Taken together, the variance among genes in apparent divergence between related species may be too much for a simple divergence model that does not include gene flow. In some cases a pattern of extreme variation among genes in levels of divergence may be so striking that a conclusion of gene exchange seems warranted even without the benefit of statistics. Such seems to be the case for the M and S forms of *Anopheles gambiae*. A genome-wide study showed that most of the genome has been homogenized through gene flow between the two forms, except for three small "islands" of differentiation, which are thought to contain genes that are important in keeping these forms distinct in sympatry (Turner & Hahn 2007, Turner et al. 2005).

This general approach, which focuses on variation in levels of divergence among genes, can be statistical if we have an understanding of how much different loci should vary in a null model that does not have any gene flow. Wakeley & Hey (1997) developed the theory for the numbers of shared and fixed differences between two populations diverging under a simple isolation model with no gene flow (**Figure 2*a***). By fitting a data set to this model, and then by comparing the actual numbers of shared and fixed differences to data simulated without gene flow under the fitted model, a statistical test of gene flow can be conducted (Wang et al. 1997). Some of the limitations of this approach include the fact that it lacks a way to estimate gene flow levels and that it cannot

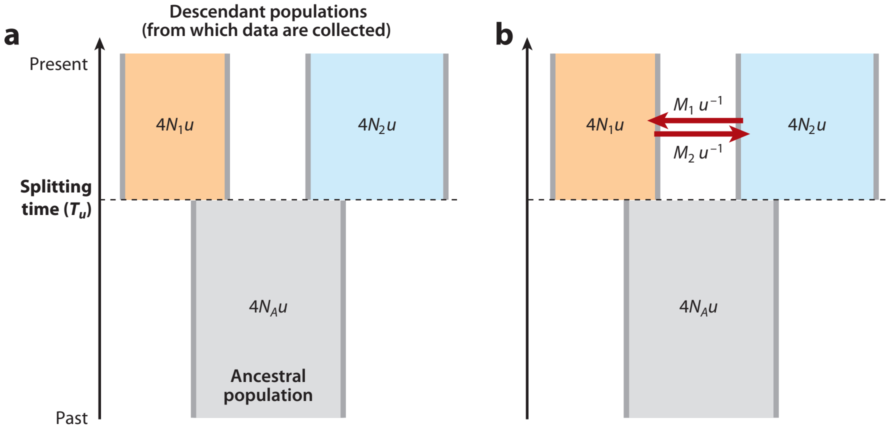

Figure 2

(*a*) The isolation model with three population size parameters and a splitting time parameter. Effective population sizes (*N*), migration rates per gene copy per generation (*M*), and time since population splitting (*T*) are all scaled by the mutation rate (*u*) of the genes being studied. Effective population sizes are assumed to be constant. The ancestral population is assumed to have been present indefinitely back into the past. (*b*) The isolation-with-migration (IM) model includes two parameters for gene flow between the sampled populations. Gene flow rates are assumed to be constant over the time period since population splitting. Migration arrows represent the movement of genes as time moves forward; alternative representations of the IM model show migration arrows in the coalescent direction backward in time.

be applied to data that departs from the infinite sites mutation model under which the theory of fixed and shared differences was developed.

Quantifying Gene Flow Using Isolation-with-Migration Models

Figure 2*b* shows the isolation-with-migration (IM) model studied by Nielsen & Wakeley (2001). It is the same as the isolation model in **Figure 2*a***; however, in addition it allows that the two daughter populations have been exchanging genes at a constant rate since splitting. This model includes six parameters, three for population size, two for gene exchange, and one splitting time term. To estimate the parameters of the IM model, Nielsen & Wakeley (2001) devised a way to approximate the likelihood:

$$P(X \mid \Theta) = \int_{\psi} P(X \mid G) P(G \mid \Theta)\, dG, \qquad 1.$$

where X represents the data (a genetic data set with samples from each of two related populations), Θ denotes the set of six parameters that are to be evaluated, and G represents a gene tree or genealogy. The calculation of $P(X \mid G)$, the probability of the data set given a particular genealogy, requires an appropriate mutation model for the genetic data, whereas the estimation of $P(G \mid \Theta)$, the probability of the genealogy given the parameters, is based on the coalescent process under the IM model. However, because the true genealogy is not known, it is necessary to consider all possible genealogies in order to fully connect the data X and the model parameters Θ; hence, there is the integration in Equation 1 over ψ, the set of possible genealogies (Felsenstein 1988, Griffiths 1989).

IM: isolation-with-migration model or analysis

The challenge of Equation 1 is that, because genealogies are complex, it is generally not possible to treat the integration over genealogies in a literal sense for purposes of calculation. The spectrum of genealogical possibilities (in terms of topology, branch lengths, and gene flow events) is so vast that a complete evaluation of Equation 1 is virtually impossible. Nielsen & Wakeley's (2001) idea was to adapt the Bayesian approach of Wilson & Balding (1998), in which both G and Θ (that is, genealogies and model parameters) are included in a Markov chain Monte Carlo (MCMC) simulation, to the IM model. This simulation yields an approximation that converges on a full likelihood-based solution. In addition, because it is likelihood-based, the method brings with it the benefits and flexibility of likelihood-based statistics, including parameter estimates and likelihood-ratio tests, and it uses all of the information in the data that bears on the model (Fisher 1922, 1925).

One of the main results from an analysis under this method is a curve for each migration parameter with an estimated posterior probability for every value of migration within the range specified by the user. The location of the peak of this curve can be taken as an estimate of the parameter. The difference between the probabilities at that peak and at a migration rate of zero can be used for a likelihood-ratio test of the null hypothesis of zero gene flow (Hey 2010, Nielsen & Wakeley 2001).

There have been several extensions and alterations of the original method of Nielsen & Wakeley (2001), including changes that permit the use of data sets with multiple loci (Hey & Nielsen 2004), relax the assumption of constant population sizes (Hey 2005), reduce the state space of MCMC simulation to just genealogies (Hey & Nielsen 2007), and allow the study of more than two related populations (Hey 2010). The methods have been implemented as various publicly available computer programs including *MDIV* (Nielsen & Wakeley 2001), *IM* (Hey 2005, Hey & Nielsen 2004), *IMa* (Hey & Nielsen 2007), and *IMa2* (Hey 2010). All of these methods share the assumption that the history of sampled populations can be reasonably represented by an IM model. Additionally, they share the following assumptions:

1. The sampled genes are under selective neutrality. This does not necessarily mean strict neutrality (e.g., a model in which deleterious mutations occur but are removed by natural selection can still be approximated by the model); however, genes that have been the target of recent selective sweeps will be in violation.
2. Mutation has followed the assumptions of the particular mutation model in use. The programs can accommodate the infinite sites mutation model (Kimura 1969), the HKY finite sites mutation model (Hasegawa et al. 1985), and the stepwise mutation model for use with microsatellites (Ohta & Kimura 1973).
3. No recombination within sampled genes.
4. Free recombination between sampled genes.

Most of the practical difficulties that arise when using these methods fall into three areas:

1. Not having sufficient data. Rarely has history been such that data from a single locus can provide resolution on all six parameters, and even multilocus data sets may not generate parameter estimates with small confidence intervals.
2. Having data that does not meet the assumptions of the model. Failures of this sort can be of a great many types.
3. Slow mixing of the Markov chain simulation. Large data sets and data from some kinds of histories usually require very long run times, and figuring out how long to run a program can be difficult. Investigators typically need to do repeated independent runs to assess convergence of their MCMC simulations.

Markov chain Monte Carlo (MCMC): a method of computer simulation to generate random values from a complex probability distribution

***IM*:** a computer program that performs an IM analysis

***IMa*:** a computer program that performs an IM analysis

Recently, other methods, which avoid some of these difficulties, have been developed for the IM model. Although they do not use all of the information in the data, because of their reliance on summary statistics, the MCMC method of Becquet & Przeworski (2007) and the Approximate Bayesian Computation (ABC) method of Lopes et al. (2009) do not require the assumption of no recombination within loci.

WHAT ISOLATION-WITH-MIGRATION ANALYSES REVEAL ABOUT GENE FLOW DURING DIVERGENCE

We conducted a meta-analysis of published research articles that used the *IM* and *IMa* programs to estimate gene flow among recently diverged taxa. We identified ∼250 papers that used either of these programs and that were in the Web of Science database as of August 4, 2009. Because we were interested in obtaining information on gene flow during divergence, we did not further consider articles about closely related populations of the same species. All studies addressing questions of divergence (either between species, subspecies, or lower level differentiation) were included as long as there was also some evidence reported in the paper of differentiation (e.g., phenotypic or ecological). We excluded several studies in which the overall inference design appeared to be circular because groups were defined based on the same genetic markers that were used to fit the IM model. There were also several cases where multiple studies (often by different researchers) addressed the divergence among the same group of taxa. In these cases, we selected one study that seemed most complete in terms of the amount of data or thoroughness of IM analysis. Finally, we kept in our survey only articles that reported quantitative parameter estimates that were relevant for subsequent analysis. The final list used for the meta-analysis includes 49 studies (see **Supplemental Table 1**; follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**).

Many of the studies reported on multiple analyses that differed, for example, in the number of loci included, or the data set that was used, or on the parameters estimated. In these cases, we chose the set of parameter estimates that most fit the particular question we addressed (see below). Many studies dealt with more than two taxa and, hence, presented multiple pairwise analyses. To avoid the nonindependence of data points from studies with $k > 2$ taxa, we randomly selected $k - 1$ of the pairwise comparisons. In cases where the researchers performed separate analyses for different kinds of loci, we randomly selected one of the data sets for further analysis.

For each of the articles in the final set, we collected all the available information regarding parameter estimates obtained using the programs mentioned above, as well as noting other relevant methodological aspects (e.g., whether or not tests of IM assumptions were performed, whether or not there was a test of nonzero gene flow, etc.). In many cases, only a subset of estimated parameters were reported. We tried to maximize the available information by performing conversions using an appropriate mutation rate (provided by the researchers, calculated based on information in the paper, or estimated from other conversions). Because we were interested in evaluating overall patterns of divergence with gene flow, we computed the mean of the locus-specific migration rates in articles in which the researchers did not report an overall migration rate. The final dataset (**Supplemental Table 1**) thus includes a mixture of estimates taken directly from the articles as well as some values based on calculations from those estimates and other values reported in the studies.

The Distribution of Estimated Population Migration Rates among Studies

To understand the impact of gene flow on divergence, we would like to know the effective number of gene migrations received by a population per generation. This value is also known as the

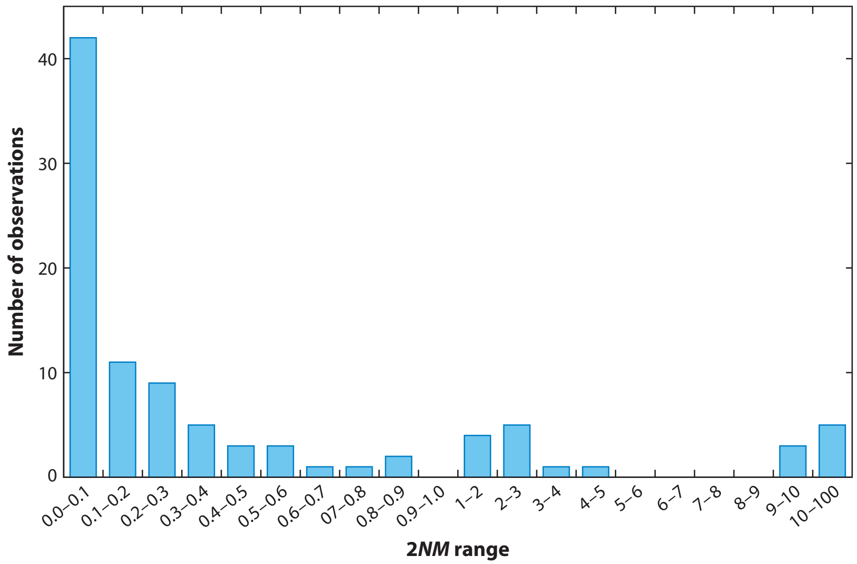

Figure 3

Counts of estimated mean 2*NM* values among studies. From those studies listed in **Supplemental Table 1** that reported 2*NM* values, or from which these values could be calculated, the mean of the values for each direction is given. Note the changing X-axis scale for larger values.

population migration rate or 2*NM* (Wright 1931). One simple way to estimate this is take the product of an estimated population size parameter (4*Nu*) and the estimated migration parameter for genes moving into that population (*M/u*) and divide by two. **Figure 3** shows a histogram for the average 2*NM* value (that is, the mean of the values for each direction between two populations) from the studies that met the criteria described above. Each of the estimated values has a variance, and any individual estimated 2*NM* that is not zero may or may not be statistically different from zero. However, the histogram in **Figure 3** may still be a useful estimate of the distribution of 2*NM* values in nature. At least two features merit notice: the sharp falloff of counts for values much above zero and the substantial fraction and long tail of gene flow estimates that are above zero.

In a neutral model without selection, values of 2*NM* greater than or equal to one can prevent populations from accumulating much divergence (Wright 1931). For studies of divergence, where disruptive selection may be acting, a finding of even a small amount of gene flow can be quite interesting. However, because IM analyses fit a neutral model, without selection, they do not reveal how much selection is acting against gene flow. Any particular value of 2*NM* may reflect the actual rate of hybrid formation and reproduction, if hybrids and their progeny have high fitness, or it may reflect only the movement of genes that manage to pass between populations despite low fitness in hybrids and their progeny. For this reason, the estimated values of 2*NM* must be treated as measures of net gene flow, because they reflect the passage of genes after some unknown amount of selection against gene flow has acted.

Gene Flow Versus Population Splitting Time

If we take as given that complete reproductive isolation comes eventually to all sister species, then we expect gene flow levels to decline for increasing times since populations began to separate.

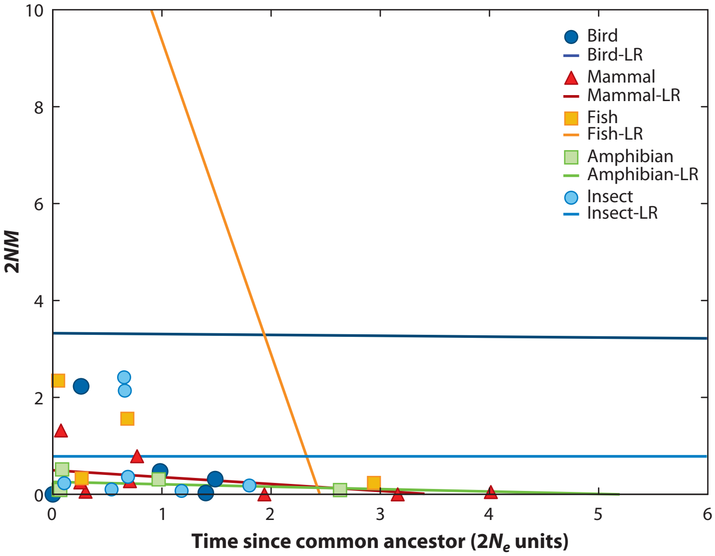

Figure 4

Mean $2NM$ against splitting time by taxonomic group. Five major taxonomic groups were represented by five or more studies. Points and linear regression (LR) trend lines are shown (a small number of points fell outside the plotted area). Splitting time is in units of $2N_e$ generations and is calculated as the estimated value of the splitting parameter, divided by one half of the mean of the estimated population size parameters for the sampled populations.

To see if this pattern is found among the surveyed studies, we calculated the mean value of $2NM$ for each pair of sampled populations and plotted these values against estimates of the population splitting time in units of $2N_e$ generations (that is, the timescale on which genetic drift leads to divergence). The overall trend is for $2NM$ values to decline with larger splitting times; however, there is a wide scatter to the points, and the slope of the linear regression is not significantly different from zero (results not shown). **Figure 4** shows points and trend lines determined by linear regression for those large taxonomic groups that are represented by more than three points. All of the slopes are negative (though the regression lines for insects and birds are very flat), so there is a suggestion of the expected relationship. However, the main impression gained from this analysis is that the process of gene flow decline over time occurs with a wide variance.

Can Other Factors Cause the Appearance of Gene Flow in an Isolation-with-Migration Analysis?

Without the addition of more parameters, there are obviously many kinds of demographic history that may be overlooked by the six-parameter IM model in **Figure 2*b***. Structure within populations, changes in the sizes of populations, and changes in the amount of gene flow over time are just some of the demographic complexities that investigators might wish to study and which are not directly accessible in a six-parameter IM model. One way to address this concern, particularly if a lot of data are available, is to use more complicated models (e.g., Fagundes et al. 2007, Hey 2005). Another general approach to this issue is to simulate data under models that violate the assumptions of the basic IM model and evaluate the performance of the programs on the simulated data. Two recent studies have examined by simulation the performance of the *IM* and *IMa* programs (Becquet &

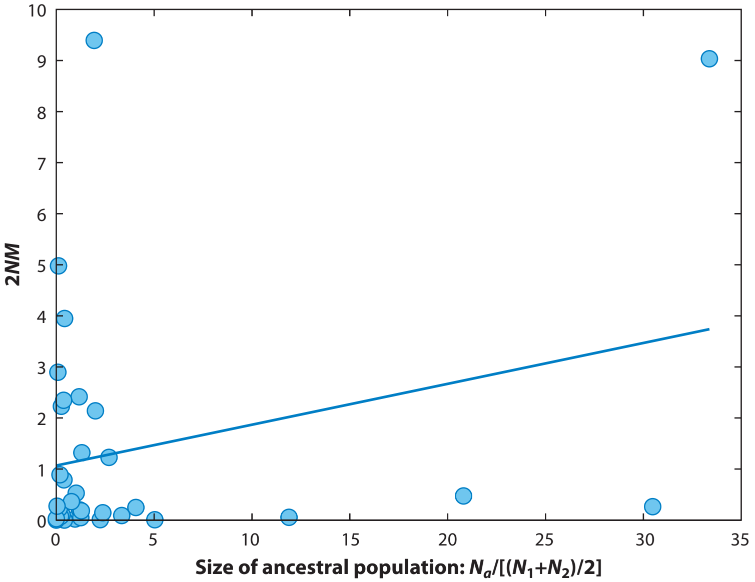

Figure 5

Mean $2NM$ plotted against the relative size of the ancestral population. The size of the ancestral population is calculated by dividing the estimated value by the mean of the estimated sizes of the descendant populations.

Przeworski 2009, Strasburg & Rieseberg 2010) in the face of violations of many of the assumptions. Not surprisingly, the ancestral population size parameter, which is the parameter that pertains to the oldest times in the model and, thus, one to which relatively little information in the data apply, was often the one most likely to be affected by departures from model assumptions.

Becquet & Przeworski (2009) identified a scenario in which a structured ancestral population becomes panmictic immediately prior to splitting into two descendant populations, which can create falsely positive signals of gene exchange after the splitting event. This kind of history is similar to one that has been discussed in which the ancestral population has exchanged genes with other nonsampled populations and ends up as a result having an estimated size that is considerably larger than the sampled populations (Won et al. 2005). The basic IM model assumes that the ancestral population, prior to splitting, had persisted in panmixia and isolation from all other populations forever. Failures of this assumption, particularly ones that add variation to the depth of the genealogies in that ancestral population, can be expected to impact estimates of parameters that apply to times more recent than the splitting event. One way to accommodate such histories is to include more than two sampled populations and, thus, more than one population splitting event (Hey 2010), although such models also require much more data.

We can partly assess whether or not the studies that show signs of nonzero gene flow do so because of added ancestral variation. If an ancestor had added variation (e.g., due to population structure or to gene exchange with another population), then its estimated size might be substantially larger than those of descendant populations (Won et al. 2005). If this same phenomenon also contributed to nonzero gene flow estimates, we might observe a positive association between these variables. In **Figure 5**, we plot the mean value of *2NM* against the size of the ancestral population as a fraction of the size of descendant populations. There is a suggestion of a positive relationship, but this is largely driven by a single point with high values for both variables, and the regression does not approach statistical significance.

FUTURE DIRECTIONS

Divergence can and does lead to new species in the presence of gene flow. However, this process is expected to arise by a complex interaction between disruptive selection, and many questions remain. At least two main areas of questions and research can be foreseen—one that is functional and a second that is historical and demographic. First, what are the details of the interaction between disruptive selection and linkage? Answering this question requires identification of the gene targets of selection and the recombinational landscape around them in the genome. Second, what has been the detailed history of gene exchange during the divergence process? In particular, we would like to know how often disruptive selection itself can be the initial trigger for disequilibrium or if most cases of divergence with gene flow began with divergence arising between separated populations. We would also like to know how gene flow rates have varied over time as well as along the chromosomes in relation to the targets of selection.

SUMMARY POINTS

1. The question of whether or not two species can diverge from one another while they share a geographic area and while they exchange genes has interested evolutionary biologists since Darwin proposed such a model as his "principle of divergence."
2. Models of divergence in sympatry are more complex than models of divergence in allopatry, but both classes of models generally require that selection have affected two or more genes, and both classes of models require features that create linkage disequilibrium between loci affected by selection.
3. Under sympatry, linkage disequilibrium can be caused directly by disruptive selection, particularly if multilocus heterozygotes have reduced fitness. It can also be facilitated by tight linkage, assortative mating, and chromosomal rearrangements that reduce crossing over in multilocus heterozygotes.
4. Statistical methods for studying exchange under an isolation-with-migration (IM) model have been widely used in recent years. This model includes population-size parameters for an ancestral and two descendant populations, as well as a splitting time term and two parameters for gene exchange.
5. A review of papers that have used the *IM* and *IMa* computer programs shows that (*a*) a plurality of studies find low or zero gene flow, while (*b*) many studies do find evidence of nonzero gene flow (sometimes at quite high levels) between diverging populations or species.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was partially supported by a research grant from the Fundação para a Ciência e a Tecnologia [FCT, Portugal: PTDC/BIA-BDE/66210(2006)] and by a postdoctoral fellowship from the FCT [SFRH/BPD/28869(2006)] to C.P. This work was also supported by a National Institutes of Health grant (R01GM078204) and a National Science Foundation grant (DEB-0949561) to J.H.

LITERATURE CITED

Bateson W. 1909. Heredity and variation in modern lights. In *Darwin and Modern Science*, ed. AC Seward, pp. 85–101. Cambridge, UK: Cambridge Univ. Press

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17:1505–19

Becquet C, Przeworski M. 2009. Learning about modes of speciation by computational approaches. *Evolution* 63:2547–62

Bolnick DI, Fitzpatrick BM. 2007. Sympatric speciation: models and empirical evidence. *Annu. Rev. Ecol. Evol. Syst.* 38:459–87

Bush G. 1994. Sympatric speciation in animals: new wine in old bottles. *Trends Ecol. Evol.* 9:285–88

Butlin RK. 2005. Recombination and speciation. *Mol. Ecol.* 14:2621–35

Carneiro M, Ferrand N, Nachman MW. 2009. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* 181:593–606

Coyne JA, Orr HA. 2004. *Speciation*. Sunderland, MA: Sinauer Assoc.

Dobzhansky T. 1937. *Genetics and the Origin of Species*. New York: Columbia Univ. Press

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, et al. 2007. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* 104:17614–19

Feder J, Roethele J, Filchak K, Niedbalski J, Romero-Severson J. 2003. Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics* 163:939–53

Felsenstein J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. *Evolution* 35:124–38

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22:521–65

Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A* 222:309–68

Fisher RA. 1925. Theory of statistical estimation. *Proc. Camb. Philos. Soc.* 22:700–25

Gavrilets S. 2003. Perspective: models of speciation: what have we learned in 40 years? *Evolution* 57:2197–215

Geraldes A, Ferrand N, Nachman MW. 2006. Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* 173:919–33

Gould SJ. 2002. *The Structure of Evolutionary Theory*. Cambridge, MA: Belknap Press of Harvard Univ. Press

Griffiths RC. 1989. Genealogical-tree probabilities in the infinitely-many-site model. *J. Math. Biol.* 27:667–80

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160–74

Hey J. 2005. On the number of new world founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol.* 3:0965–75

Hey J. 2010. Isolation with migration models for more than two populations. *Mol. Biol. Evol.* 27:905–20

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–60

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. USA* 104:2785–90

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61:893–903

Kirkpatrick M, Ravigné V. 2002. Speciation by natural and sexual selection: models and experiments. *Am. Nat.* 159:S22–35

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* 5:288–98

Lopes JS, Balding D, Beaumont MA. 2009. PopABC: a program to infer historical demographic parameters. *Bioinformatics* 25:2747–49

Machado CA, Kliman RM, Markert JM, Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and its close relatives. *Mol. Biol. Evol.* 19:472–88

Muller HJ. 1940. Bearings of the *Drosophila* work on systematics. In *The New Systematics*, ed. J Huxley, pp. 185–268. Oxford, UK: Clarendon Press

Navarro A, Barton NH. 2003. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution* 57:447–59

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. *Genetics* 158:885–96

Noor MA, Grams KL, Bertucci A, Almendarez Y, Reiland JA, Smith KR. 2001a. The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. *Evolution* 55:512–21

Noor MA, Grams KL, Bertucci LA, Reiland J. 2001b. Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci. USA* 98:12084–88

Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res. Camb.* 22:201–4

Orr HA. 1991. Is single-gene speciation possible? *Evolution* 45:764–69

Orr HA. 1996. Dobzhansky, Bateson, and the genetics of speciation. *Genetics* 144:1331–35

Rice WR, Hostert EF. 1993. Laboratory experiments on speciation: what have we learned in 40 years. *Evolution* 47:1637–53

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16:351–58

Rieseberg LH, Whitton J, Gardner K. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* 152:713–27

Salzburger W, Meyer A. 2004. The species flocks of East African cichlid fishes: recent advances in molecular phylogenetics and population genetics. *Naturwissenschaften* 91:227–90

Slotman MA, Reimer LJ, Thiemann T, Dolo G, Fondjo E, Lanzaro GC. 2006. Reduced recombination rate and genetic differentiation between the M and S forms of *Anopheles gambiae* s.s. *Genetics* 174:2081–93

Strasburg JL, Rieseberg LH. 2010. How robust are "isolation with migration" analyses to violations of the IM model? A simulation study. *Mol. Biol. Evol.* 27:297–310

Stump AD, Fitzpatrick MC, Lobo NF, Traoré S, Sagnon N, et al. 2005. Centromere-proximal differentiation and speciation in *Anopheles gambiae*. *Proc. Natl. Acad. Sci. USA* 102:15930–35

Tammone W. 1995. Competition, the division of labor, and Darwin's principle of divergence. *J. Hist. Biol.* 28:109–31

Trickett AJ, Butlin RK. 1994. Recombination suppressors and the evolution of new species. *Heredity* 73:339–45

Turner TL, Hahn MW. 2007. Locus- and population-specific selection and differentiation between incipient species of *Anopheles gambiae*. *Mol. Biol. Evol.* 24:2132–38

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3:e285

Via S. 2001. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol. Evol.* 16:381–90

Wagner M. 1873. *The Darwinian Theory and the Law of the Migration of Organisms*. Transl. JL Laird. London: E. Stanford (From German)

Wakeley J, Hey J. 1997. Estimating ancestral population parameters. *Genetics* 145:847–55

Wang RL, Wakeley J, Hey J. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* 147:1091–106

Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. *Genetics* 150:499–510

Won YJ, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. *Proc. Natl. Acad. Sci. USA* 102:6581–86

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159

Wright S. 1951. The genetical structure of populations. *Ann. Eugen.* 15:323–54

Wu CI. 2001. The genic view of the process of speciation. *J. Evol. Biol.* 14:851–65

Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH. 2007. Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics* 175:1883–93