MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2387–2397 doi: 10.1111/mec.13610

GENOMICS OF HYBRIDIZATION

Powerful methods for detecting introgressed regions from population genomic data

BENJAMIN K. ROSENZWEIG,* JAMES B. PEASE,*† NORA J. BESANSKY‡ § and MATTHEW W. HAHN*¶

*School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA, †Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA, ‡Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA, §Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556, USA, ¶Department of Biology, Indiana University, Bloomington, IN 47405, USA

Abstract

Understanding the types and functions of genes that are able to cross species boundaries—and those that are not—is an important step in understanding the forces maintaining species as largely independent lineages across the remainder of the genome. With large next-generation sequencing data sets we are now able to ask whether introgression has occurred across the genome, and multiple methods have been proposed to detect the signature of such events. Here, we introduce a new summary statistic that can be used to test for introgression, $RND_{\min}$, that makes use of the minimum pairwise sequence distance between two population samples relative to divergence to an outgroup. We find that our method offers a modest increase in power over other, related tests, but that all such tests have high power to detect introgressed loci when migration is recent and strong. $RND_{\min}$ is robust to variation in the mutation rate, and remains reliable even when estimates of the divergence time between sister species are inaccurate. We apply $RND_{\min}$ to population genomic data from the African mosquitoes *Anopheles quadriannulatus* and *A. arabiensis*, identifying three novel candidate regions for introgression. Interestingly, one of the introgressed loci is on the X chromosome, but outside of an inversion separating these two species. Our results suggest that significant, but rare, sharing of alleles is occurring between species that diverged more than 1 million years ago, and that application of these methods to additional systems are likely to reveal similar results.

Keywords: adaptation, bioinfomatics, genomics, population genetics, speciation

Received 16 October 2015; revision received 17 February 2016; accepted 22 February 2016

Introduction

The increasing availability of whole-genome sequencing data has shed new light on speciation and the genomic patterns of divergence between closely related lineages (e.g. Kulathinal *et al.* 2009; Renaut *et al.* 2012; Cui *et al.* 2013; Martin *et al.* 2013; Brandvain *et al.* 2014; Brawand *et al.* 2014; Carneiro *et al.* 2014; Jónsson *et al.* 2014; Lamichhaney *et al.* 2015). This work has supported the view that diverging populations can hybridize after considerable periods of time apart, and has shown that traces of introgression via secondary contact can be found in the genomes of diverse taxa. For instance, Neandertal alleles are found in modern European humans (Green *et al.* 2010), and introgression of colour-pattern genes between species of *Heliconius* butterflies may have played a role in an adaptive radiation (The *Heliconius* Genome Consortium 2012). Many other instances of introgression have been found, possibly with adaptive consequences (e.g. Song *et al.* 2011; Brand *et al.* 2013; Norris *et al.* 2015; reviewed in Hedrick 2013). Understanding the types and functions of gene that are able to cross species boundaries, as well as those that are not introgressed, is an important step in understanding the forces maintaining species as largely

Correspondence: Benjamin K. Rosenzweig, Fax: (812) 855-6705;
E-mail: bkrosenz@indiana.edu and Matthew W. Hahn,
Fax: (812) 855-6705; E-mail: mwh@indiana.edu

independent lineages across the remainder of the genome (Seehausen *et al.* 2014). Therefore, the development of methods that can accurately identify when introgression is taking place—and the precise regions that are introgressed—is now an active area of research.

The genetic signatures of introgression are not always readily apparent, and can be masked by a number of factors (cf. Cruickshank & Hahn 2014). This is especially true when introgression is rare across the genome, and exists only in small "islands of introgression" (Garrigan *et al.* 2012). In these cases the goal is to identify these islands accurately, and methods have been developed to do this using both an approximate Bayesian computation framework (Roux *et al.* 2013) or a fully Bayesian framework (Sousa *et al.* 2013). However, model-based methods may fail because multiple aspects of the underlying model are violated in real data. For detecting introgression between sister species, multiple summary statistic methods have been developed (e.g. Joly *et al.* 2009; Geneva *et al.* 2015) all of which depend on the notion that introgressed regions will show higher similarity between species than nonintrogressed regions. These methods can also be prone to false positives, and may not be sensitive to all introgression events. Because introgressed regions should have higher sequence similarity than background (nonintrogressed) regions, regions of lower mutation rate can mimic introgressed regions. Conversely, introgression soon after a speciation event may be difficult to identify because there is extensive sharing of alleles due to incomplete lineage sorting (ILS). Finally, if the introgression was recent or of low magnitude, individuals sampled from the recipient taxon simply may not carry an introgressed lineage, or only a small fraction of individuals will carry one. An ideal measure of introgression should be robust and powerful under a variety of such conditions.

In this article, we introduce a new test for introgression and apply it to population genomic data from mosquitoes in the genus *Anopheles*: the sister species, *A. quadriannulatus* and *A. arabiensis* (Fontaine *et al.* 2015). Our test is a natural extension of other recent methods (Joly *et al.* 2009; Geneva *et al.* 2015), and is based on the minimum sequence divergence found between haplotypes in sister taxa. We find that our new statistic offers a modest increase in power over others that use population genetic data, and that it is robust to multiple violations that may confound inferences using other tests.

Materials and methods

We consider the situation in which we want to identify introgressed regions between sister lineages. Such cases exclude the use of powerful methods that can only be used with three or more lineages because they are based on the different topologies produced by hybridization (i.e. the "ABBA-BABA" test and related *D*- and *F*-statistics; Huson *et al.* 2005; Reich *et al.* 2009; Green *et al.* 2010; Durand *et al.* 2011; Patterson *et al.* 2012; Liu *et al.* 2014; Pease & Hahn 2015). The data collected may consist of a single sequence from each species or multiple sequences from each species, with or without an outgroup (assumed to have no possibility of introgression). We can also consider both phased and unphased data, though the statistic we propose requires phased haplotypes.

Given multiple sequences from within two species, the most well-known method for identifying introgressed loci is the fixation index, F_{ST} (Wright 1931). F_{ST} does not require an outgroup and can be calculated from single SNPs or other markers; it does not require full sequence data. If full sequence data is obtained F_{ST} does not require phased data, though there are several related statistics that do use phased haplotypes (e.g. Hudson *et al.* 1992). F_{ST} quantifies the normalized difference in allele frequencies between populations, and exceptionally low values of F_{ST} at a locus relative to the background are a good indicator of gene flow.

However, several factors can confound inferences of migration using F_{ST}, including natural selection (Charlesworth 1998; Noor & Bennett 2009; Cruickshank & Hahn 2014). An alternative measure that is largely robust to the effects of linked selection is d_{XY}. If we define $d_{x,y}$ as number of sequence differences between any two sequences, x and y, in two taxa, X and Y (divided by the number of sites compared), then d_{XY} is the average distance between all sequences in the two species. Low values of d_{XY} indicate recent introgression. Note that calculations of d_{XY} do not require the sequences to be phased, and could use only one sequence from each species (often this is denoted simply as d). Neither d_{XY} nor F_{ST} are very sensitive to the presence of low-frequency migrants (Geneva *et al.* 2015). This means that analyses using these statistics alone will fail to detect recent introgression (e.g. Murray & Hare 2006).

To detect even rare introgressed lineages, Joly *et al.* (2009) proposed using the minimum sequence distance between any pair of haplotypes from two taxa. Defining $d_{x,y}$ as above, the minimum sequence distance, $d_{\min}$, is $\min_{x \in X, y \in Y}\{d_{x,y}\}$, the minimum distance among all pairings of haplotypes in the two species. The logic behind this method is that any two sequences that are highly similar to each other—and therefore represent a coalescence more recent than the population divergence time—can only be explained by introgression (Fig. 1). By comparing the observed $d_{\min}$ to the expected values under a no-migration model, we can obtain positive evidence for introgression when the observed value is

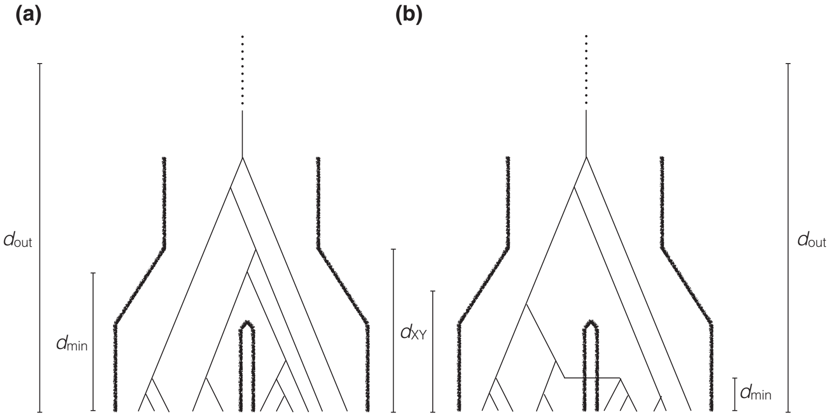

Fig. 1 Explanation of the statistics used here. In both panels two species are represented, with multiple sampled sequences per species. A representative gene genealogy is shown, and divergence to the outgroup is represented by a dotted line. (a) A history with no migration. Here, $d_{\min}$ is equal to the distance between the closest two lineages coalescing before the speciation event. (b) A history with migration. Here $d_{\min}$ is equal to the distance between lineages related via introgression. The average value of d_{XY} also goes down slightly in this case, but importantly d_{out} is no different than in panel a.

in the lower tail of this distribution (below some specified *P*-value). The null distribution of $d_{\min}$ when there is no migration is generally produced via coalescent simulations, but could simply be a comparison among genomic regions if the expectation is that most did not introgress. This method has high power when its assumptions are met (Joly *et al.* 2009), but it makes a number of assumptions that are likely to be violated quite often.

The most important assumption made by both d_{XY} and $d_{\min}$ is that there is no variation in the neutral mutation rate among loci. Variation among loci with low neutral mutation rates can be mistaken for a locus that has experienced a recent introgression event, unless this variation is explicitly included in the simulated null model. Multiple solutions have been proposed to account for mutation rate variation. One such alternative (that does not require haplotypes) is to account for the relative node depth (RND) of the two taxa compared to an outgroup (Feder *et al.* 2005). RND is defined as the quotient of d_{XY} between the two species to the average distance from each to an outgroup:

$$RND = \frac{d_{XY}}{d_{out}}$$

where $d_{out} = (d_{XO} + d_{YO})/2$, and d_{XO} is the average distance between species X and the outgroup, O, while d_{YO} the average distance between species Y and the outgroup. Low substitution rates are reflected in shortened branch lengths both between X and Y and between each of them and the outgroup. Therefore, *RND* is robust to low mutation rates as long as mutation rates have been constant across the tree. However, *RND* is still not sensitive to low-frequency migrants.

Geneva *et al.* (2015) introduced a method that is relatively sensitive to recent migration, while still being robust to variation in mutation rates. Their test statistic, $G_{\min}$, is defined as:

$$G_{\min}\frac{d_{\min}}{d_{XY}}$$

where $d_{\min}$ and d_{XY} are the same as above. Because a lower mutation rate is expected to affect all haplotypes equally, the normalization by the average distance between all haplotypes in the two species will account for variable rates of evolution among loci. While $G_{\min}$ is robust to variable mutation rates, Geneva *et al.* (2015) report low power (<0.5) across the range of parameters they tested. In fact, $G_{\min}$ was only reported to have any ability to detect introgression when both the migration probability and relative migration time are low (Figure S1 in Geneva *et al.* 2015). This is likely due to the fact that as migrant lineages rise in frequency, d_{XY} also gets lower. As a migrant haplotype approaches fixation, the ratio of $d_{\min}$ to d_{XY} approaches 1.

To develop a statistic that is both robust to mutation rate variation and sensitive to low-frequency migrants, we propose to combine the best aspects of $d_{\min}$, $G_{\min}$, and *RND*. Here we introduce $RND_{\min,}$ defined as:

$$RND_{\min} = \frac{d_{\min}}{d_{out}}$$

Low substitution rates are reflected in shortened branch lengths to the outgroup, so $RND_{\min}$ (like *RND*) is robust to variable mutation rates. Similarly, like both $d_{\min}$ and $G_{\min}$, $RND_{\min}$ should be sensitive to even rare migrant haplotypes. In addition, we expect $RND_{\min}$ to be powerful even when migrants are high in frequency.

We therefore expect that it will have higher power than other methods.

Simulations

We used simulations to investigate the statistical properties of multiple statistics used to detect introgression, under a variety of migration scenarios. All coalescent simulations were carried out using the coalescent simulator, msmove (Garrigan & Geneva 2014), a variant of ms (Hudson 2002). Artificial sequences were generated with Seq-Gen (Rambaut & Grassly 1997). Statistical analysis of simulated data as well as the *Anopheles* data was conducted with the MVFtools software library (Pease & Rosenzweig 2015).

We use p_M to denote the migration probability (the fraction of a population composed of migrants), τ_D the divergence time of the ancestral populations (measured in units of effective population size, $4N_e$) and τ_M the time of the migration event relative to the divergence time (also in units of $4N_e$; migration events are assumed to occur instantaneously). We assume a sample size of $n = 15$ haploid individuals from each population, and simulate 10 000 loci of 1000 bp in length from each combination of values in $\tau_D \in \{1/4, 2/4, \ldots, 8\}$, $\tau_M \in \{\tau_D/20, 2\tau_D/20, \ldots, 19\tau_D/20\}$, and $p_M \in \{0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, \ldots, 0.8, 0.85, 0.9\}$. The divergence time of the outgroup was fixed at $40N_e$ generations before present.

We also generated a null distribution of values ($n = 200\,000$ loci) for each divergence time from simulations with no migration ($p_M = 0$). We consider a locus being tested as "positive" for migration (i.e. significant) if its test statistic falls in the bottom 5% of the null distribution of test statistics. When the power of a test is 1 it indicates that every locus simulated with a history of migration was detected by our test statistic of choice.

Anopheles data

We used a population genomic data set generated by Fontaine *et al.* (2015). The data set consisted of samples from *A. quadriannulatus* ($n = 10$ diploid individuals) and *A. arabiensis* ($n = 12$ diploid individuals), with the reference genome of *A. christyi* as the outgroup. Each set of samples from a species was sequenced using Illumina short-read technology, with mean read-depth of 12.6X for *A. quadriannulatus* and 14.8X for *A. arabiensis*. Reads from each species were aligned to the reference genome created for that species (Neafsey *et al.* 2015), with alignments of each reference genome to the *A. gambiae* reference used to create a multiple sequence alignment. Further details of sample collection and sequencing can be found in Fontaine *et al.* (2015). Because $RND_{\min}$ requires haplotypic data, samples were computationally phased using BEAGLE (Browning & Browning 2007). We calculated $RND_{\min}$ in 50 kb nonoverlapping windows along all five chromosomal arms, with minimum thresholds for sequence alignment length and number of samples per species. To identify significant windows, a null distribution was simulated with a divergence time of $256N_e$ generations. This value takes the divergence time estimate of 1.28 million years (Fontaine *et al.* 2015) and converts it to coalescent units by assuming 10 generations per year (Fontaine *et al.* 2015) and $N_e = 50\,000$. The estimate of the effective population size comes from assuming that the expected value of π is $4N_e\mu$, coupled with the observation that π is approximately 0.001 in both species (Fontaine *et al.* 2015) and per-generation per-site mutation rates are 5×10^{-9} in *Drosophila melanogaster* (Schrider *et al.* 2013). To be conservative, we also calculated the null distribution using simulations with a divergence time of $12N_e$ generations (that is, assuming $N_e = 1\,000\,000$). In both cases average d_{out} and d_{XY} were comparable to the values observed in the *Anopheles* data set.

Results

Statistical power of tests for introgression

We simulated a wide range of histories of introgression, from just after the initial species split ($\tau_M = 0.90$) to long after this split ($\tau_M = 0.05$), and including both very high ($p_M = 0.9$) and very low ($p_M = 0.1$) probabilities of migration. For all of these histories we calculated F_{ST} (using the formula given in Geneva *et al.* 2015), d_{XY}, $G_{\min}$, and our new statistic, $RND_{\min}$, asking in each case whether a locus had a significant signature of introgression for each statistic separately. Because we did not simulate loci with mutation-rate heterogeneity, d_{XY} and *RND* should have exactly the same power, as should $RND_{\min}$ and $d_{\min}$. Therefore, we do not present separate results for *RND* and $d_{\min}$.

Figure 2 reports the power of F_{ST}, d_{XY}, $G_{\min}$ and $RND_{\min}$ to detect introgression (results on false positives are presented in the next section). As expected, all of the statistics have the highest power when introgression has been recent and strong: all three statistics detect ~100% of loci with a history of introgression in this area of parameter space (the upper left-hand corner of Fig. 2a–d). When introgression follows closely on the heels of divergence all three statistics have less power, likely because not enough sequence differences have accumulated since the split to distinguish introgressed loci from the loci in the null distribution. Likewise, when the probability of migration is lower—which is equivalent to there being only a small fraction of the

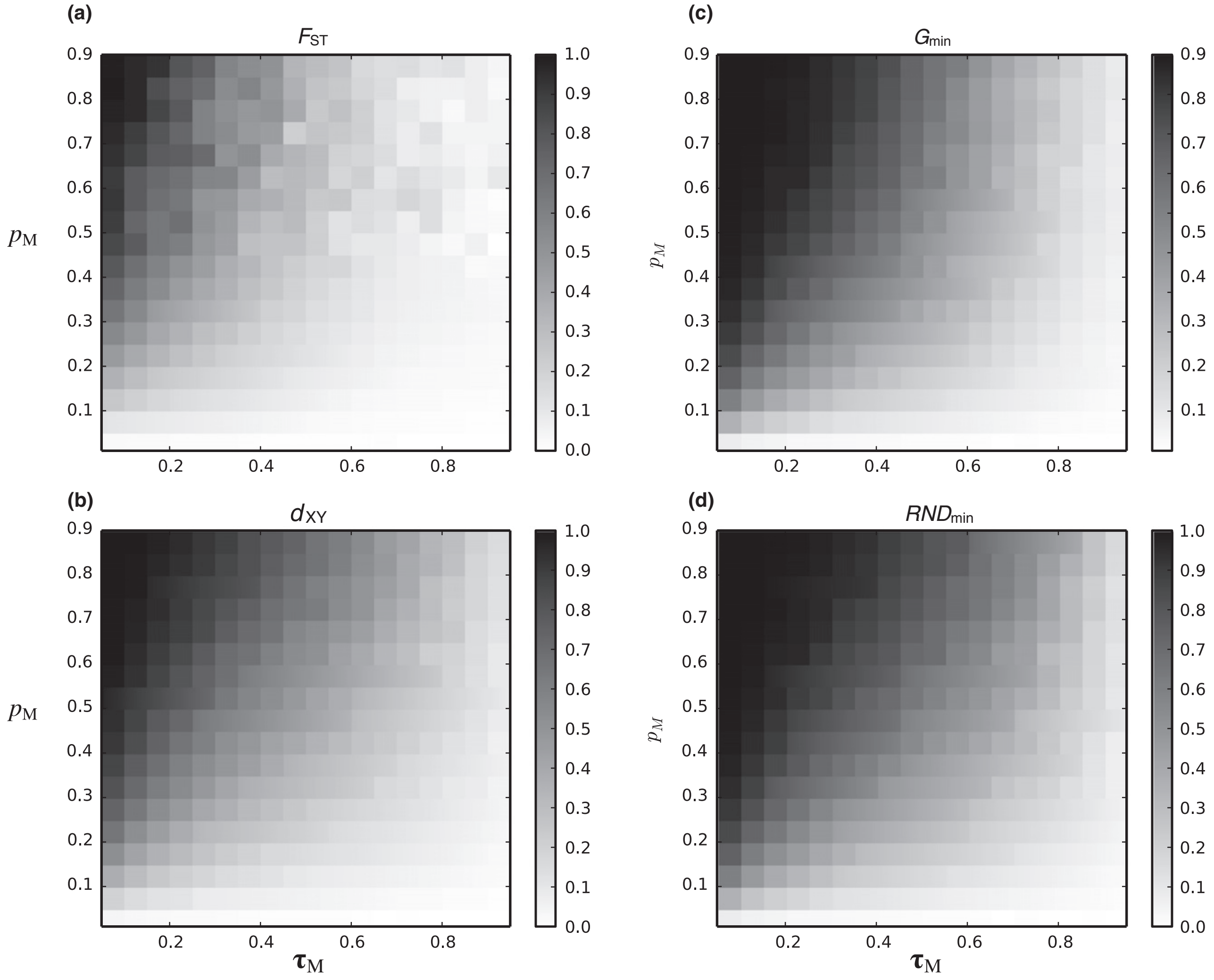

Fig. 2 Comparison of power to detect introgression for four statistics. For all panels, the history simulated was one of species divergence at time $\tau_D = 4N_e$ generations ago, across multiple values of the relative time (τ_M) and level (p_M) of migration. Also for all panels, the power of each test in every square on the grid is given by the colour scale shown to the right. (a) Power of F_{ST} to detect migration. (b) Power of d_{XY} to detect migration. (c) Power of G_{min} to detect migration. (d) Power of RND_{min} to detect migration.

population composed of migrants at the moment gene flow occurs—it becomes harder to detect introgressed loci. This effect is almost completely determined by whether a sample contains a migrant lineage at a locus: we used msmove to track the presence of migrant lineages, finding that our ability to identify a locus as introgressed was highly correlated with the presence of at least one migrant lineage in the population sampled ($r = 0.73$; $P < 10^{-320}$). When migration probabilities are low, or even when they are at moderate levels but not in the recent past, samples simply do not contain the descendants of migrants. In these cases it is impossible to detect the footprint of introgression.

Also as expected, the relative power of the three statistics was $RND_{min} > G_{min} > d_{XY} > F_{ST}$. We measured this as the total sum of power values calculated for each of the 306 parameter combinations tested (17 values of τ_M × 18 values of p_M); this can be seen visually as the fraction of darker squares in Fig. 2a–d. The measure of the mean sequence divergence between populations, d_{XY}, had the most power when migrant lineages are common, but cannot detect the signatures of introgression contained within rare migrant lineages (Fig. 2b). G_{min} does better at detecting low-frequency migrant lineages, just as it was designed to do (Fig. 2c). The new statistic introduced here, RND_{min}, does better still, especially increasing power as τ_M becomes higher (Fig. 2d). The difference between RND_{min} and G_{min} can be seen more clearly in Fig. 3, which shows the difference in power between the two statistics. While the two measures have similar power over much of parameter space, $RND_{\min}$ gains the most power when migration events occurred close to the divergence time, possibly because many or most of the sampled lineages have migrant ancestors.

Our results suggest that all of these tests for introgression have relatively high power. This is in contrast with the results reported for $G_{\min}$ in Geneva *et al.* (2015), where the power of this statistic was never greater than 0.5, and then only for an extremely limited portion of parameter space where both τ_M and p_M were <0.1 (Figure S1 in Geneva *et al.* 2015). The difference in results is entirely due to the different ways used to calculate power. Geneva *et al.* (2015) calculated power (referred to as sensitivity in that paper) as the proportion of simulated loci with migrant lineages with a value of $G_{\min}$ in the tail of the distribution of all values of $G_{\min}$ from simulations with migration. That is, they did not identify outliers by referring to a simulated null distribution without migration, but instead used the distribution of the same $G_{\min}$ values with migration (via a *Z*-test). This necessarily limits the amount of power this test can have, though if one assumes introgression has been rare enough in the genome (such that all migrant lineages lie in the 5% tail of the sample distribution) the method works well. Note that if we calculate power in the same way, we reproduce their results for $G_{\min}$ exactly (data not shown).

One concern with any method that depends on a simulated null distribution is the accuracy of such simulations. The dependence on τ_D can sometimes be alleviated by coestimating divergence times and gene flow (e.g. Melo-Ferreira *et al.* 2012), but such approaches are not practical with whole-genome data sets. In all previous results we have calculated power using null simulations carried out with the correct value of τ_D. To test the sensitivity of our results to the match between our simulations and the true values, in the next section we examine the behaviour of $RND_{\min}$ when τ_D is misspecified.

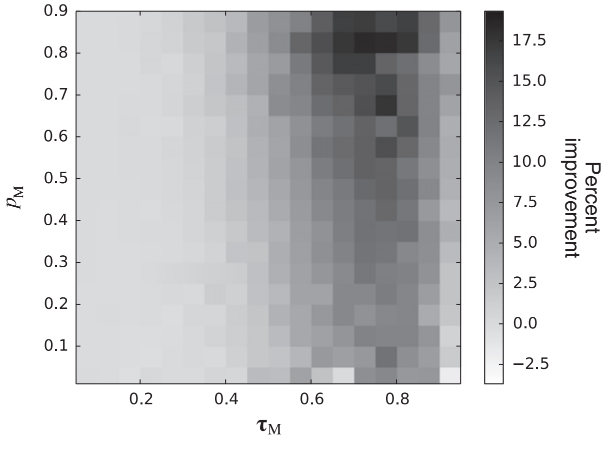

Fig. 3 Relative improvement in power to detect introgression using $RND_{\min}$ over $G_{\min}$. The percent improvement in each cell is given by the colour scale to the right. Values represent the results from Fig. 2d relative to Fig. 2c.

Sensitivity of results to accurate null simulations

We simulated two species with a divergence time of $\tau_D = 4N_e$ generations ago experiencing an introgression event of magnitude $p_M = 0.6$, at time $\tau_M = 0.3$ relative to divergence (i.e. $1.2N_e$ generations ago). If we simulate the null distribution of these two species experiencing no migration and diverging (accurately) at $\tau_D = 4N_e$, the power of $RND_{\min}$ is 86% (Fig. 2d). However, researchers will not always know the correct divergence time, and so we should not expect to achieve this level of power in all real settings.

In fact, we predict that our power will almost certainly go down when using data collected from nature. This is because divergence times are estimated from the same data that are being tested for introgression: if introgression is actually occurring, then sequence divergence among our samples will be lower than that expected given the species split time, and τ_D will be underestimated. Since the value of τ_D used for the null simulation is now lower than the true value, our power will go down—this is equivalent to increasing τ_M (i.e. moving from left to right across any panel in Fig. 2). To demonstrate this effect, we calculated the power to detect introgression using $RND_{\min}$ when the value of τ_D used for simulation varies. Figure 4a demonstrates that, as predicted, underestimating τ_D leads to a loss of power and a reduction in the false positive rate (see next paragraph). In the extreme, when the value of τ_D used in the simulation is the same as the time of introgression ($\tau_M = 0.3$), we have no power to detect introgression. This is because migrant lineages are not any more similar in terms of sequence divergence than simulated nonmigrant lineages.

When τ_D is higher than the true value, we see no loss in power. This is because the observation of migrant lineages is just as surprising at the true value of τ_D as it is at higher values. After reaching its theoretical maximum near $\tau_D = 8N_e$ (when the true value is $4N_e$), the power plateaus. However, this raises the concern that simulating values of τ_D that are too high could increase the false positive rate when there is no migration actually occurring. To investigate this effect we again simulated two species with a divergence time of $\tau_D = 4N_e$ generations ago, but with no introgression; this is equivalent to the null simulations used in Fig. 2. We then asked what fraction of these simulated loci lie in the lower tail of the distribution of additional null sim-

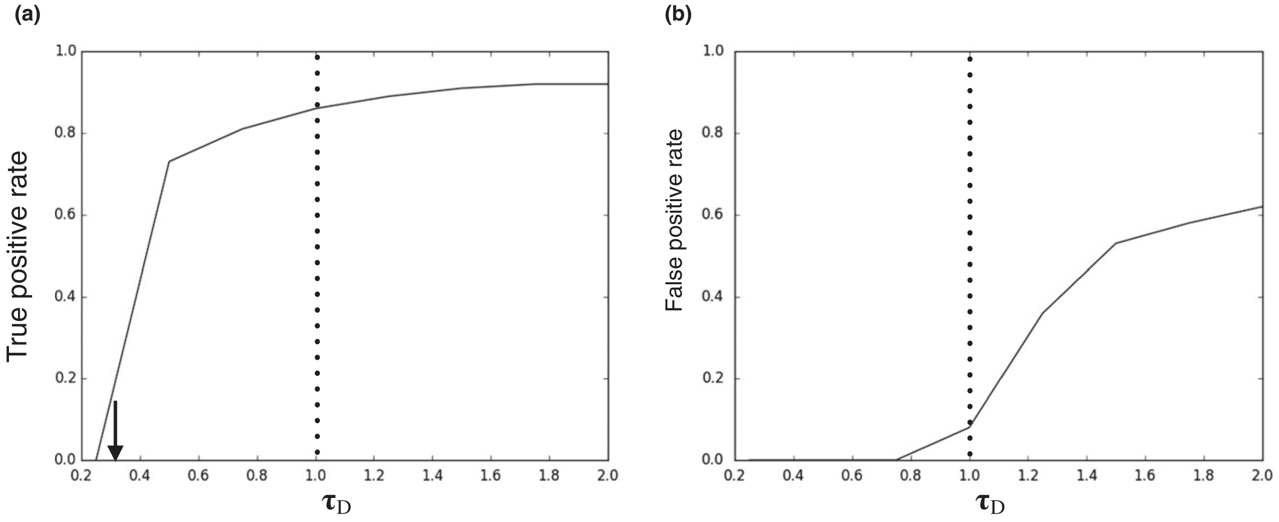

Fig. 4 Power and robustness of RND_{min} when τ_D is misspecified. (a) A history with migration. The two lineages split at time $\tau_D = 1$ (vertical dotted line), measured in units of $4N_e$ generations, and migration occurred at time $\tau_M = 0.3$ (arrow). For null simulations conducted across multiple values in $\tau_D \in \{0.25, 0.5,\ldots,2\}$, the power to detect true positives is reported. (b) A history with no migration. The two lineages again split at time $\tau_D = 1$, and again null simulations are conducted across multiple values in $\tau_D \in \{0.25, 0.5,\ldots,2\}$. In this case no migration occurred, and we therefore report the rate at which false positives were reported.

ulations with no introgression, but with varying values of τ_D. Figure 4b shows that when τ_D is underestimated, there is no increase in false positives. Conversely, when τ_D is overestimated there is an increase in false positives because observed sequences are more similar than expected. While this is a general concern, in practice we believe that researchers are more likely to under- than over-estimate τ_D when there is any gene flow whatsoever. If there is no gene flow, a conservative approach would be to use the lower-bound of whatever estimate of τ_D has been made.

Introgressive hybridization in Anopheles

We obtained whole-genome sequence data from multiple individuals in each of the sister species *A. arabiensis* and *A. quadriannulatus* (Fontaine *et al.* 2015). All of the *A. quadriannulatus* individuals were collected in Zimbabwe, while the *A. arabiensis* individuals were collected in Tanzania, Cameroon, and Burkina Faso (Table S4 in Fontaine *et al.* 2015). These sampling locations reflect the much smaller species range of *A. quadriannulatus*, which is found only in southern Africa. We looked for signals of recent introgression using our new statistic, RND_{min}, calculated in 50-kb windows across all five chromosome arms in the *Anopheles* genome (Fig. 5).

The average value of d_{XY} between *A. arabiensis* and *A. quadriannulatus* was 0.0397 and the average distance to *A. christyi* (i.e. d_{out}) was 0.2263. The average of RND_{min} was 0.154 across the genome, with values of 0.155 and 0.145 on the autosomes and X, respectively. The X value is within the range of the individual autosomal averages (range 2R = 0.141 to 2L = 0.168), which is as expected since dividing by divergence to the outgroup controls for chromosome-specific differences in substitution rates. The range of RND_{min} among individual 50-kb windows was from 0.013 to 0.362, which is much larger than that of our null simulations (0.078–0.304 assuming $256N_e$ generations since divergence and 0.065–0.222 assuming $12N_e$ generations). Also as expected (Hudson & Coyne 2002), there is little incomplete lineage sorting after even $12N_e$ generations, so we have good power to detect outliers given the estimated divergence time between these species.

We detected three main regions with low outlier values of RND_{min} from the middle of chromosome arms: on chromosome 3L from 23–24 Mb, on the X at 13.85 Mb, and on 3R at 32.5 Mb (Fig. 5). All three of these regions show significantly reduced values of RND_{min} using cut-offs generated by either of our null simulations ($P < 0.001$ for both), and are therefore strongly suggestive of recent introgression. We were more cautious in our interpretation of low values near the centromeres and telomeres, as these regions can show low values of divergence because of linked selection in the ancestral population of *A. arabiensis* and *A. quadriannulatus* (Begun *et al.* 2007; Cruickshank & Hahn 2014). Unfortunately, comparisons with the outgroup do not appear to control as well for these reductions, possibly because the effect of reduced divergence in low-recombination regions becomes smaller

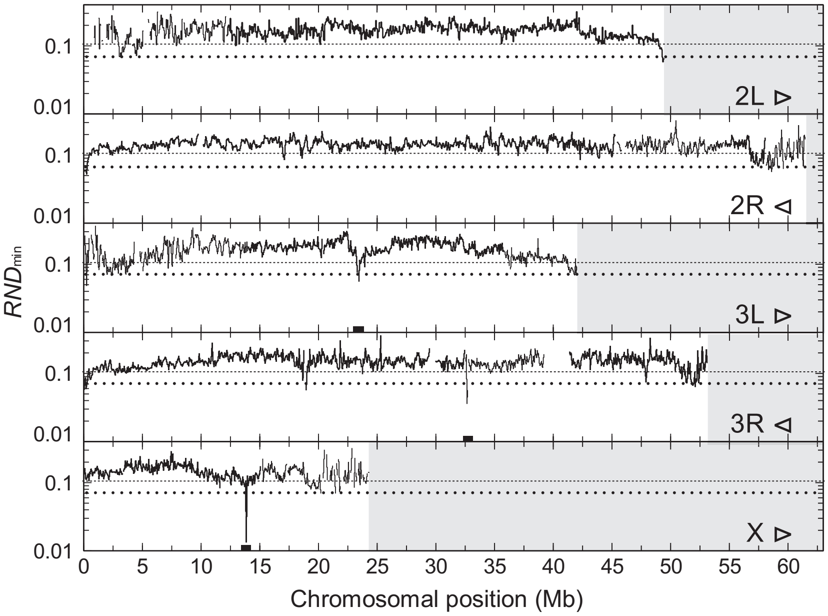

Fig. 5 Values of RND_{min} across chromosome arms in comparisons of *A. arabiensis* and *A. quadriannulatus*. For each chromosome arm values are reported for nonoverlapping 50 kb windows. The upper horizontal dotted line represents the 5% cut-off for null simulations carried out with $\tau_D = 256N_e$ generations; the lower horizontal line the 5% cut-off with $\tau_D = 12N_e$ generations. Arrows for each arm indicate the direction of the telomere, and small black bars mark the position of the three introgressions discussed in the main text. The Xag inversion extends from position 0 to 15 Mb along the X chromosome. Note the log-scale on the *y*-axis.

with longer divergence times to the outgroup. In addition, the pericentromeric and peritelomeric regions of the *Anopheles gambiae* reference genome have lower quality assemblies, and therefore we expect more variability in such regions simply due to this factor. There were few genes with annotated functions in our candidate introgressed loci, largely because of the dearth of annotated genes in the *Anopheles* genome. One of the only annotated genes was inside the X chromosome "introgression island": *septin 4* is a nucleotide binding protein that regulates cytoskeletal organization; this might point to a possible role in adaptation to the environment. Further work will have to be done to characterize the functions of other genes in these regions before inferences about the types of genes that introgress across species boundaries can be made.

Discussion

Detecting regions of the genome introgressed between taxa has become an important task, not least because of the growing realization that these may be more common than previously believed (Mallet *et al.* 2016) and that they may often be adaptive (Hedrick 2013). Because introgression homogenizes sequences between species, detecting such regions against a background of nonintrogressing DNA appears to be a straightforward task. However, due to variation in mutation rates, recent split times between species, and low-frequency migrant haplotypes, many tests can lack power or can be nonspecific. Our aim in this paper was to introduce a new statistic that will have high power but that will also lead to few false positives.

The "minimum sequence distance" family of statistics considered here provide a powerful complement to the *D*- and *F*-statistics now in common use for detecting introgression (Huson *et al.* 2005; Reich *et al.* 2009; Green *et al.* 2010; Durand *et al.* 2011; Patterson *et al.* 2012; Pease & Hahn 2015). While *D*- and *F*-statistics are appropriate for trios (or quartets) of related taxa, they rely on underlying discordant topologies to detect introgression. The statistics discussed here (d_{min}, G_{min}, and RND_{min}) can detect introgression even between sister species, but of course could be used among any pair of species in larger clades (e.g. Geneva *et al.* 2015). Furthermore, these statistics can identify instances of complete replacement by introgressed alleles. Other allele-frequency methods such as the *F*-statistics of Reich *et al.* (2009), in contrast, are effective only when introgressed alleles remain polymorphic in all populations. At the moment none of these statistics (except those introduced by Pease & Hahn 2015) can determine the direction of introgression.

Unlike methods based on topologies, pairwise methods require coalescent simulations to generate a null expectation, and their type I and type II error rates are strongly dependent on the choice of τ_D in these simulations. Care must be taken, therefore, that accurate divergence times are estimated for the taxa under consideration. These simulations also make many simplifying assumptions about demographic histories and heterogeneity in recombination rates; further work will be needed to explore how such complexities might affect statistical properties of these tests. Furthermore, these methods require phased haplotypes to accurately calculate sequence distances, an extra experimental (or computational) step not required by D- and F-statistics, or by allele-frequency-based methods such as F_{ST}.

All minimum-distance methods will fail to detect introgression in cases where the introgressed sequence is highly conserved. No matter how many samples are considered, the minimum distance will be indistinguishable from the expectation under the null model. This is also the reason that all such methods have the lowest power when introgression occurs soon after the species split: there has been little time for the accumulation of nucleotide substitutions, and introgressed regions will not stand out from the background. Conversely, without correcting for the average substitution rate in a region, regions with low rates can appear to be introgressed when larger numbers of substitutions have accumulated across the genome, even when no introgression has occurred. This is the logic behind the original *RND* statistic (Feder *et al.* 2005), which normalizes the sequence divergence between pairs of species potentially exchanging genes using divergence to an outgroup. Although this method has most often been used to find regions that are not introgressing when gene exchange seems widespread (e.g. Nachman & Payseur 2012; Carneiro *et al.* 2014), here we have used it to ensure the robustness of our statistic for identifying regions that are introgressing.

We found that RND_{min} had high power to detect introgression across much of the parameter space examined (Fig. 2c). Surprisingly, we also found that G_{min} had high power (Fig. 2b), contrary with what had been reported by the authors of the paper describing this statistic (Geneva *et al.* 2015). This apparent difference is due to the way that power was calculated in the two papers (see Results for details); all findings are consistent if we assume that introgression is very rare and calculate power in the same manner as Geneva *et al.* (2015). We believe that the calculations carried out here are more biologically relevant to the task of detecting introgression. While RND_{min} does offer a modest increase in power over G_{min} (Fig. 3), we were also surprised that it did not do better in comparison. Our expectation was that G_{min} would not be able to detect high-frequency migrant lineages, as these would lower both d_{min} and d_{XY} and consequently G_{min} would approach 1. An explanation for the ability of G_{min} to detect such cases lies in the observation that at short divergence times (i.e. low τ_D)—even without introgression—this statistic can be much less than 1 (see Fig. 2 in Geneva *et al.* 2015). If we imagine the case where the migrant lineage becomes fixed at a locus, then the value of G_{min} at this locus will be equivalent to a divergence time of τ_M because this is the time at which all lineages between the species last shared a common ancestor. And at low τ_D, G_{min} will be appreciably lower than 1, as stated above. Therefore, even this statistic has more power than anticipated to detect high-frequency migrant lineages. We conclude that there is likely to be no noticeable different between G_{min} and RND_{min} applied to data from nature.

Using single reference genomes and D-statistics, Fontaine *et al.* (2015) identified three major introgression events among lineages in the *Anopheles gambiae* species complex (see also Wen *et al.* in press to this special issue). While one of these events involved *A. quadriannulatus* and two involved *A. arabiensis*, no gene flow was reported between these two species. However, D-statistics cannot detect introgression between sister species, so it is not surprising that none was detected between *A. quadriannulatus* and *A. arabiensis*. Despite the fact that *A. quadriannulatus* is only found in southeastern Africa (the samples used here were collected in Zimbabwe), the much larger distribution of *A. arabiensis* fully overlaps this range (Lanzaro & Lee 2013; the samples used here came from Tanzania, Cameroon, and Burkina Faso). There are reports of low levels of hybridization in nature between these species (<0.1%; White 1974; Coluzzi *et al.* 1979), and *A. arabiensis* appears to have recently hybridized with other species in the complex (Weetman *et al.* 2014), so finding introgression is certainly plausible. Using RND_{min}, we were able to identify at least three good candidates for introgressed regions. Though there are no genes with recognizably interesting functions in these regions, one important observation is that there is evidence for introgression on the X chromosome. *A. quadriannulatus* and *A. arabiensis* differ on the X by a compound inversion, denoted *Xbcd*, but the location of the introgressed window we identified lies outside (distal) to the breakpoints of this inversion (Fontaine *et al.* 2015). Therefore, much like the pattern previously found between *A. arabiensis* and the ancestor of *A. gambiae* (which also differ in the presence of the *Xag* inversion; Fontaine *et al.* 2015), there is evidence for introgression on the X only outside inverted regions. This further supports the role of inversions in reducing gene flow between species (Noor *et al.* 2001; Rieseberg 2001).

The ability to identify the regions and genes introgressing (and not introgressing) between species has been a major contribution of genomics to the study of hybridization. Through these genes, the hope is that we will begin to understand the evolutionary forces that promote or inhibit the sharing of genes among species, and therefore be able to better understand the processes promoting and inhibiting species divergence.

Acknowledgements

We thank Dan Garrigan and Anthony Geneva for helpful discussions about their work, and Scott Emrich and Michael Fontaine for assistance with the mosquito data. This work was supported by National Institutes of Health grant R01-AI076584 to NJB and MWH.

References

Begun DJ, Holloway AK, Stephens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Brand CL, Kingan SB, Wu L, Garrigan D (2013) A selective sweep across species boundaries in *Drosophila*. *Molecular Biology and Evolution*, **30**, 2177–2186.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics*, **10**, e1004410.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, **513**, 375–381.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**, 1084–1097.

Carneiro M, Albert FW, Afonso S *et al.* (2014) The genomic architecture of population divergence between subspecies of the European rabbit. *PLoS Genetics*, **10**, e1003519.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.

Coluzzi M, Sabatini A, Petrarca V, Di Deco MA (1979) Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **73**, 483–497.

Cruickshank TC, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Cui R, Schumer M, Kruesi K *et al.* (2013) Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. *Evolution*, **67**, 2166–2179.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**, 2239–2252.

Feder JL, Xie X, Rull J *et al.* (2005) Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. *Proceedings of the National Academy of Sciences*, **102**, 6573–6580.

Fontaine MC, Pease JB, Steele A *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**, 1258524.

Garrigan D, Geneva AJ (2014) msmove, Available from http://dx.doi.org/10.6084/m6089.figshare.1060474.

Garrigan D, Kingan SB, Geneva AJ *et al.* (2012) Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Research*, **22**, 1499–1511.

Geneva AJ, Muirhead CA, Kingan SB, Garrigan D (2015) A new method to scan genomes for introgression in a secondary contact model. *PLoS ONE*, **10**, e0118621.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Hedrick PW (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, **22**, 4606–4618.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution*, **56**, 1557–1565.

Hudson RR, Slatkin M, Maddison W (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

Huson DH, Klöpper T, Lockhart PJ, Steel MA (2005) Reconstruction of reticulate networks from gene trees. In: *Research in Computational Molecular Biology* (eds Miyano S, Satoru M, Jill M, Simon K, Sorin I, Pavel AP, Michael W), pp. 233–249. Springer, Berlin.

Joly S, McLenachan PA, Lockhart PJ (2009) A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, **174**, E54–E70.

Jónsson H, Schubert M, Seguin-Orlando A *et al.* (2014) Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proceedings of the National Academy of Sciences*, **111**, 18655–18660.

Kulathinal RJ, Stevison LS, Noor MA (2009) The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genetics*, **5**, e1000550.

Lamichhaney S, Berglund J, Almen MS *et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, **518**, 371–375.

Lanzaro GC, Lee Y (2013) Speciation in Anopheles gambiae—The distribution of genetic polymorphism and patterns of reproductive isolation among natural populations. In: *Anopheles Mosquitoes: New Insights into Malaria Vector* (ed Manguin S). InTech, Rijeka, Croatia.

Liu KJ, Dai J, Truong K *et al.* (2014) An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, **10**, e1003649.

Mallet J, Besansky N, Hahn MW (2016) How reticulated are species? *BioEssays*, **38**, 140–149.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, **23**, 1817–1828.

Melo-Ferreira J, Boursot P, Carneiro M *et al.* (2012) Recurrent introgression of mitochondrial DNA among hares (*Lepus* spp.) revealed by species-tree inference and coalescent simulations. *Systematic Biology*, **61**, 367–381.

Murray MC, Hare MP (2006) A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. *Molecular Ecology*, **15**, 4229–4242.

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 409–421.

Neafsey DE, Waterhouse RM, Abai MR *et al.* (2015) Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science*, **347**, 1258522.

Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, **98**, 12084–12088.

Norris LC, Main BJ, Lee Y *et al.* (2015) Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proceedings of the National Academy of Sciences*, **112**, 815–820.

Patterson N, Moorjani P, Luo Y *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Pease JB, Hahn MW (2015) Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, **64**, 651–662.

Pease JB, Rosenzweig BK (2015) Encoding data using biological principles: the Multisample Variant Format for phylogenomics and population genomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi:10.1109/tcbb.2015.2509997.

Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, **13**(3), 235–238.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature*, **461**, 489–494.

Renaut S, Maillet N, Normandeau E *et al.* (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 354–363.

Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**, 351–358.

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. *Molecular Biology and Evolution*, **30**, 1574–1587.

Schrider DR, Houle D, Lynch M, Hahn MW (2013) Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, **194**, 937–954.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Song Y, Endepols S, Klemann N *et al.* (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between Old World mice. *Current Biology*, **21**, 1296–1301.

Sousa VC, Carneiro M, Ferrand N, Hey J (2013) Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics*, **194**, 211–233.

The Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Weetman D, Steen K, Rippon EJ *et al.* (2014) Contemporary gene flow between wild *An. gambiae s.s.* and *An. arabiensis*. *Parasites & Vectors*, **7**, 345.

Wen D, Yu Y, Hahn MW, Nakhleh L (2016) Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, **25**, 2361–2372.

White GB (1974) *Anopheles gambiae* complex and disease transmission in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **68**, 278–298.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

B.K.R. and M.W.H. designed the study. B.K.R. and J.B.P. conducted the analyses. B.K.R., N.J.B., and M.W.H. wrote the article.

Data accessibility

Aligned DNA sequences in MVF format are deposited in the Dryad repository: doi:10.5061/dryad.f7 h13. Software is available in the MVFTools github repository at https://github.com/jbpease/mvftools.