Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*

Daniel R. Schrider[*,†,1], Julien Ayroles[§], Daniel R. Matute[‡], Andrew D. Kern[*,†]

[*]Department of Genetics, and [†]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, New Jersey 08554

[§]Ecology and Evolutionary Biology Department; Lewis Sigler Institute for Integrative Genomic, Princeton University, Princeton, New Jersey, 08540

[‡]Biology Department, University of North Carolina, Chapel Hill, NC 27510.

[1]Corresponding author: Department of Genetics, Rutgers University, 604 Allison Rd., Piscataway, NJ 08854. E-mail: dan.schrider@rutgers.edu

ABSTRACT

Hybridization and gene flow between species appears to be common. Even though it is clear that hybridization is widespread across all surveyed taxonomic groups, the magnitude and consequences of introgression are still largely unknown. Thus it is crucial to develop the statistical machinery required to uncover which genomic regions have recently acquired haplotypes via introgression from a sister population. We developed a novel machine learning framework, called FILET (Finding Introgressed Loci via Extra-Trees) capable of revealing genomic introgression with far greater power than competing methods. FILET works by combining information from a number of population genetic summary statistics, including several new statistics that we introduce, that capture patterns of variation across two populations. We show that FILET is able to identify loci that have experienced gene flow between related species with high accuracy, and in most situations can correctly infer which population was the donor and which was the recipient. Here we describe a data set of outbred diploid *Drosophila sechellia* genomes, and combine them with data from *D. simulans* to examine recent introgression between these species using FILET. Although we find that these populations may have split more recently than previously appreciated, FILET confirms that there has indeed been appreciable recent introgression (some of which might have been adaptive) between these species, and reveals that this gene flow is primarily in the direction of *D. simulans* to *D. sechellia*.

INTRODUCTION

Up to 10% of animal species have the ability to hybridize with other species (Mallet 2005). The ubiquity of hybridization upon secondary contact of diverging populations has led to the study of hybrid zones and the phenotypic consequences of hybridization (Barton and Hewitt 1985). Whole-genome sequencing has confirmed the notion that introgression, the genetic exchange between species through fertile hybrids, is also common between closely related species (Begun et al. 2007; Kulathinal et al. 2009; Martin et al. 2013; Brandvain et al. 2014; Fontaine et al. 2015) and in some instances between divergent species (Nürnberger et al. 2016; Turissini and Matute 2017). This is perhaps best known from the case of Neanderthal hybridization with non-African human populations (Green et al. 2010; Sankararaman et al. 2014), which has left modern human genomes with clear examples of introgressed Neanderthal alleles. Depending on the genetic architecture of reproductive isolation (i.e., number of hybrid incompatibilities, dominance of those incompatibilities), introgression might be deleterious (True et al. 1996; Harris and Nielsen 2016; Juric et al. 2016). Those loci that contribute to reproductive isolation, and as such to the persistence of species in the face of hybridization, should be less likely to be introgressed (Turner et al. 2005). On the other hand, much of the genome may be porous to introgression between closely related species if the net effect of such introgression is fitness neutral. Thus if we could reliably delineate those regions of the genome that have and have not experienced introgression among species, and the magnitude of selection against them, we may be able to understand the genetic underpinnings of reproductive isolation.

Genetic exchange between populations can also provide a source of adaptive alleles that may facilitate adaptation to new environments (reviewed in Hedrick 2013). Rather than waiting for one or more new beneficial mutations to arise, a species faced with a new environment may be able to receive these alleles via gene flow from a sympatric species already adapted for that environment (e.g. if the donor population migrated to this new environment first and/or adapted to it more rapidly). For instance, adaptation to high altitude in Tibetans appears to have been caused by introgression of alleles from an archaic Denisovan-like source into modern humans (Huerta-Sánchez et al. 2014). Another particularly well-studied system of adaptive introgression comes from *Heliconius* butterflies where gene exchange has facilitated the origin and maintenance of mimetic rings (Pardo-Diaz et al. 2012) and even of hybrid species (Melo et al. 2009; Salazar et al. 2010). Clearly, hybridization and introgression play an important role in the origin or demise of new species. Yet these isolated examples are not sufficient to elucidate the importance of introgression a source of genetic variation. A reliable framework for the inference of introgressed alleles is therefore sorely needed.

Recent work on uncovering introgressed loci has focused on the use of population genomic data from pairs of species of distinct populations. Largely the methods devised have consisted of new summary statistics that capture elements of the expected coalescent genealogy under a model of recent introgression between species. For example, values of the F_{ST} statistic will be lower in the presence of gene flow (e.g. Neafsey et al. 2010). Another popular point of

departure has been the d_{xy} statistic of Nei and Li (1979) which measures the average pairwise distance between alleles sampled from two populations. Joly et al. (2009) modified this approach by taking the minimum rather than the mean of these pairwise divergence values, termed d_{min}. d_{min} is thus sensitive to abnormally short branch lengths between alleles drawn from two populations, as would be expected under a model of recent introgression. Similarly, Geneva et al. (2015) and Rosenzweig et al. (2016) devised with their own statistics to detect introgression, both based on d_{min} but with added robustness to variation in the neutral mutation rate. Each of these statistics has attractive properties and adequate power in some instances, however no one statistic has perfect sensitivity in every scenario.

In order to fill this void, we introduce a new method for finding introgressed loci based on supervised machine learning that we call FILET (Finding Introgressed Loci using Extra Trees Classifiers). FILET combines a large number of summary statistics (Materials and Methods) that provide complementary information about the shape of the genealogy underlying a region of the genome. These summary statistics include both previously developed statistics (including, but not limited to, those based on d_{min} and d_{xy}) as well as 5 new summary statistics that we describe below. Our reasoning for this approach was that by combining many statistics for detecting introgression we should achieve sensitivity to introgression across a larger range of scenarios than accessible to any individual statistic. Buoyed by our recent work showing the power and flexibility of Extra Trees classifiers (Geurts et al. 2006) for population genomic inference (Schrider and Kern 2016; Schrider and Kern 2017), we leveraged this machine learning paradigm for identification of introgression. Using simulations we show that FILET is far more powerful and versatile than competing methods for identifying introgressed loci. Further we apply FILET to examine patterns of introgression between *Drosophila simulans* and its island endemic sister taxon *Drosophila sechellia*.

The speciation event that gave rise to the island endemic *Drosophila sechellia* from a *Drosophila simulans*-like ancestor is a textbook example of a specialist species that evolved from a presumably generalist ancestor (Jones 1998, 2005). Indeed, *D. sechellia* has quite remarkably specialized to breed on the toxic fruit of *Morinda citrifolia* (Louis and David 1986), while *D. simulans* (and *D. mauritiana*) do not tolerate the organic volatile compounds in the ripe fruit (Legal et al. 1994; Farine et al. 1996; Legal et al. 1999). The genetic and neurological underpinnings of this key ecological difference have been identified, at least in part (Dekker et al. 2006; Matsuo et al. 2007; Hungate et al. 2013; Huang and Erezyilmaz 2015; Shiao et al. 2015; Andrade López et al. 2017) making the *D. simulans*/*D. sechellia* pair one of the most successful cases of genetical dissection the causes of an ecologically relevant trait. While this is so, the population genetics of divergence between these species has only been examined in the context of either population samples from a handful of loci (Hey and Kliman 1993; Kliman et al. 2000; Kern et al. 2004; Legrand et al. 2009) or in the absence of population data (Garrigan et al. 2012). These studies estimated population divergence time between *D. simulans* and *D. sechellia* to be as early as ~250,000 years ago (Garrigan et al. 2012) or as old as ~413,000 years ago (Kliman et al. 2000). All population genomic surveys demonstrate that *D. sechellia* harbors little

genetic variation in comparison to *D. simulans*, perhaps as a result of a population size crash/founder event from which the population has not recovered (Hey and Kliman 1993; Legrand et al. 2009). Moreover it has been suggested that what little variation there is in *D. sechellia* shows little population genetic structure among separate island populations in the Seychelles archipelago (Legrand et al. 2009). Lastly there is some evidence of introgression between each pair of species within the *D. simulans* complex (Garrigan et al. 2012), and *D. simulans* and *D. sechellia* have been found to hybridize in the field (Matute and Ayroles 2014). Here we characterize the population genetics of divergence between *D. sechellia* and *D. simulans,* combining existing whole-genome sequences from a mainland population of *D. simulans* (Rogers et al. 2014) with newly generated genome sequences from *D. sechellia.* Applying FILET to these data confirms previous reports of introgression between these species and reveals that this gene flow is primarily in the direction of *D. simulans* to *D. sechellia.* Finally, the success of our approach underscores the potential power of supervised machine learning for evolutionary and population genetic inference.

MATERIALS AND METHODS

Statistics capturing the population genetic signature of introgression

We set out to assemble a set of statistics that could be used in concert to reliably determine whether a given genomic window had experienced recent gene flow. Several statistics that have been designed to this end ask whether there is a pair of samples exhibiting a lower than expected degree of sequence divergence within the window of interest. The most basic of these is d_{min}, the minimum pairwise divergence across all cross-population comparisons (Figure S1; Joly et al. 2009). The reasoning behind d_{min} is that even if only a single sampled individual contains an introgressed haplotype, d_{min} should be lower than expected and the introgression event may be detectable. A related statistic is G_{min}, which is equal to d_{min}/d_{xy} (Geneva et al. 2015); the presence of this term in the denominator is meant to control for variation in the neutral mutation rate across the genome. RND_{min} accomplishes this by dividing d_{min} by the average divergence of all sequences from either species to an outgroup sequence (Rosenzweig et al. 2016). The name of this statistic is derived from its constituent parts, d_{min}, and *RND* (Feder et al. 2005).

As described in the following section, we incorporated a number of previously devised statistics into our classification approach, including some of those based on d_{min}. We also included some novel statistics that we designed to have improved sensitivity to particularly recent introgression. The first of these is defined as:

$$d_{d1} = d_{min}/\pi_1$$

where π_1 is nucleotide diversity (Nei and Li 1979) in population 1. Similarly, $d_{d2} = d_{min}/\pi_2$. d_{d1} and d_{d2} statistics are so named because they compare d_{min} to diversity within populations 1 and 2, respectively. The rationale behind these statistics is that, if there has been recent introgression from population 1 into population 2, and at least one sampled chromosome from population 2

contains the introgressed haplotype, then the cross-population pair of individuals yielding the value of d_{min} should both trace their ancestry to population 1. Thus, the sequence divergence between these two individuals should on average be equal to π_1. Similarly, if there was introgression in the reverse direction d_{min} would be on the order of π_2. Following similar rationale, we devised two related statistics: $d_{d\text{-}Rank1}$ and $d_{d\text{-}Rank2}$. $d_{d\text{-}Rank1}$ is the percentile ranking of d_{min} among all pairwise divergences *within* population 1; the value of this statistic should be lower when there has been introgression from population 1 into population 2. $d_{d\text{-}Rank2}$ is the analogous statistic for introgression from population 2 into population 1. We also included a statistic comparing average linkage disequilibrium within populations to average LD within the global population (i.e. lumping all individuals from both species together), as follows:

$$Z_X = (Z_{nS1} + Z_{nS2}) / (2 \times Z_{nSG})$$

where Z_{nS1}, and Z_{nS2} measure average LD (Kelly 1997) between all pairs of variants within the window in population 1 and population 2, respectively, and Z_{nSG} which measures LD within the global population. The reasoning behind this statistic is based on the assumption that, in the presence of gene flow, LD may be elevated within the recipient population(s) but not in the global population. Figure S2 shows that the distributions of these statistics do indeed differ substantially between genealogies with and without introgression (simulation scenarios described below), especially when this introgression occurred recently. In addition to these and other statistics summarizing diversity across the two population samples, we also incorporated several single-population statistics into our classifier (see below), as these may also contain information about recent introgression. For example, separate measures of nucleotide diversity in our two population samples would contain useful information because introgression is expected to increase diversity in the recipient population, especially if the source population was large or if the two populations split long ago.

Description of FILET classifier

We used a supervised machine learning approach to assign a genomic window to one of three distinct classes on the basis of a "feature vector" consisting of a number of statistics summarizing patterns of variation within the window from two closely related populations. These three classes are: introgression from population 1 into population 2, introgression from population 2 into population 1, and the absence of introgression. Specifically, we used an Extra-Trees classifier (Geurts et al. 2006), which is an extension of random forests (Breiman 2001), an ensemble learning technique that creates a large ensemble of semi-randomly generated binary decision trees (Quinlan 1986), before taking a vote among these decision trees in order to decide which class label should be assigned to a given data instance (i.e. genomic window in our case). In an Extra-Trees classifier, the tree building process is even more randomized than in typical random forests: in addition to selecting a random subset of features when generating a tree, the separating threshold for each feature is randomly chosen, rather than selected the threshold that optimally separates the data classes. We require example regions for each class in order to train

the Extra-Trees classifier, so we used coalescent simulations to generate these training examples (described below). Our ultimate goal was to detect introgression within 10 kb windows in *Drosophila*, so to train our classifier properly we simulated chromosomal regions approximating this length (simulation details are given below). The target window size could easily be altered by changing the length of the regions simulated for training (i.e. by adjusting the recombination and mutation rates, θ and ρ).

FILET's feature vector contains a number of single-population summaries of per-base pair genetic variation: π, the variance in pairwise diversity, the density of segregating sites, the density of polymorphisms private to the population, Fay and Wu's H and θ_H statistics (Fay and Wu 2000), and Tajima's D (Tajima 1989). The feature vector also includes two single-population summary statistics that are not normalized per base pair: Z_{nS} (which is averaged across all pairs of SNPs), and the number of distinct haplotypes observed in the window. Each feature vector included values of these 9 statistics for each population, yielding 18 single-population statistics in total. In addition, the two-population statistics included in FILET's feature vector were as follows: F_{ST} (following Hudson et al. 1992), Hudson's S_{nn} (Hudson 2000), per-bp d_{xy}, per-bp d_{min}, G_{min}, d_{d1}, d_{d2}, $d_{d\text{-}Rank1}$, $d_{d\text{-}Rank2}$, Z_X, IBS_{MaxB} (the length of the maximum stretch of identity by state [IBS] among all pairwise between-population comparisons), and IBS_{Mean1} and IBS_{Mean2} which capture the average IBS tract length when comparing all pairs of sequences within populations 1 and 2, respectively. These IBS statistics are calculated by examining all pairs of individual sequences within a population (or across populations in the case of IBS_{MaxB}), noting the positions of differences, and examining the distribution of lengths between these positions (as well as between the first position and the beginning of the window and between the last position and the end of the window). Note that we did not include RND_{min} or other measures such as Patterson's D and F_4 statistics (Patterson et al. 2012) that require alignment to one or more additional species along with the focal pair, because we designed FILET so that it would not require outgroup information. We note however that through its use of supervised machine learning, FILET could easily be extended to incorporate such data. Instead, in order to improve robustness to mutational variation, we adopted the approach of drawing the mutation rate from a wide range of values when generating training examples to train FILET to classify data from our *Drosophila* samples (see below). All code necessary to run the FILET classifier (including calculating summary statistics on both simulated and real data sets, training, and classification) along with the full results of our application to *D. simulans* and *D. sechellia* (described below) are available at https://github.com/kern-lab/FILET/.

Simulated test scenarios

Following Rosenzweig et al. (2016), we used the coalescent simulator msmove (https://github.com/geneva/msmove) to simulate data for testing FILET's power to detect introgression in populations with four different values of T_D (the time since divergence): $0.25\times4N$, $1\times4N$, $4\times4N$, and $16\times4N$ generations ago, where N is the population size. For each of

these simulations the population size was held constant (i.e. the ancestral population size equals that of both daughter populations). We developed a classifier for each of these scenarios of population divergence. Supervised machine learning techniques such as the Extra-Trees classifier require training data consisting of examples from each of the three classes, but in practice a large number of example loci known to have experienced introgression may not be available. We therefore simulated training data sets for each of the four values of T_D. Again following Rosenzweig et al. (2016), the relevant parameters for each of these simulations include: T_M, the time since the introgression event, which we drew from $\{0.01\times T_D, 0.05\times T_D, 0.1\times T_D, 0.15\times T_D, \ldots, 0.9\times T_D\}$ (i.e. multiples of $0.05\times T_D$ up to 0.9, and also including $0.01\times T_D$); and P_M, the probability that a given lineage would migrate from the source population to the sink population during the introgression event, which we drew from $\{0.05, 0.1, 0.15, \ldots, 0.95\}$. We simulated an equal number of training examples for each combination of these two parameter values for both directions of gene flow, yielding 10^4 simulations in total for both of these classes, conditioning that each of these instances must have contained at least one migrant lineage. Finally, we simulated an equivalent number of samples without introgression, yielding a balanced training set (10^4 examples for each class).

Next, we computed feature vectors as described above for each of these training examples, and proceeded with training our Extra-Trees classifiers by conducting a grid search of all training parameters precisely as described in Schrider and Kern (2016), and setting the number of trees in the resulting ensemble to 100. All training and classification with the Extra-Trees classifier was performed using the scikit-learn Python library (http://scikit-learn.org; Pedregosa et al. 2011). We also calculated feature importance and rankings thereof by training an Extra-Trees classifier of 500 decision trees on the same training data (using scikit-learn's defaults for all other learning parameters), and then using this classifier's "feature_importances_" attribute. Briefly, this feature importance score is the average reduction in Gini impurity contributed by a feature across all trees in the forest, always weighted by the probability of any given data instance reaching the feature's node as estimated on the training data (Breiman et al. 1984); this measure thus captures both how well a feature separates data into different classes and how often the feature is given the opportunity to split (i.e. how often it is visited in the forest). The values of these scores are then normalized across all features such that they sum to one.

For each T_D, we evaluated the appropriate classifier against a larger set of 10^4 simulations generated for each parameter combination along a grid of values of T_M and P_M. The values of P_M were drawn from the same set as those in training as described above, while one additional possible value of T_M was included: $0.001\times T_D$. Also note that for these simulations we did not require at least one migrant lineage as we had done for training (information that is recorded by msmove). For test simulations with bidirectional migration, we did not require each replicate sample to contain at least one migrant lineage, though we modified msmove to record which samples did in fact experience migration. In addition to test examples for each direction of gene flow, we simulated 10^4 examples where no migration occurred in order to assess false positive

281 rates. In order to examine the performance of FILET when an unsampled ghost lineage was the
282 source of introgression, we generated test simulations with the same values of T_D, T_M, and P_M as
283 above, but where the source of the introgression was a third, unsampled population that separated
284 from the two sampled populations at time T_D. In all of our simulations, both for training and
285 testing, we set locus-wide population mutation and recombination rates θ and ρ to 50 and 250,
286 respectively, similar to autosomal values in 10 kb windows in *D. melanogaster* (Chan et al.
287 2012) and sampled 15 individuals from each population. We also experimented with different
288 window sizes, simulating training and test data (1,000 replicates for each class for each set) with
289 window sizes corresponding to 1 kb, 10 kb, 5 kb, and 50 kb by multiplying θ and ρ by the
290 appropriate scalar. When testing the sensitivity of our method on these data, we considered a
291 window to be introgressed if FILET's posterior probability of the no-introgression class was
292 <0.05, except for the scenario with T_D equal to $16\times4N$ generations ago in which case we used a
293 posterior probability cutoff of 0.01, as we found that this step mitigated the elevated false
294 positive rate under this scenario (reducing the rate from >10% to the estimate of 6% shown in
295 Figure S3). In windows labeled as introgressed, the direction of gene flow was determined by
296 asking which of the two introgression classes had a higher posterior probability. Note that we
297 used the same demographic scenario for both the training and test data for each T_D, and discuss
298 the implications of demographic model misspecification in the Results and Discussion.
299 In order to compute receiver operator characteristic (ROC) curves we constructed
300 balanced binary training sets composed of 10^4 examples with no introgression, and 10^4 examples
301 allowing for introgression (with equal representation to each combination of T_M, P_M, and
302 direction of introgression. The score that we obtained for each test example in order to compute
303 the ROC curve was one minus the posterior probability of no introgression as generated by the
304 Extra-Trees classifier (i.e. the classifier's estimated probability of introgression, regardless of
305 directionality).
306
307 ***Drosophila sechellia* collection**
308
309 *Drosophila sechellia* flies were collected in the islands of Praslin, La Digue, Marianne and Mahé
310 with nets over fresh *Morinda* fruit on the ground. All flies were collected in January of 2012.
311 Flies were aspirated from the nets by mouth (1135A Aspirator – BioQuip; Rancho Domingo,
312 CA) and transferred to empty glass vials with wet paper balls (to provide humidity) where they
313 remained for a period of up to three hours. Flies were then lightly anesthetized using FlyNap
314 (Carolina Biological Supply Company, Burlington, NC) and sorted by sex. Females from the
315 *melanogaster* species subgroup were individualized in plastic vials with instant potato food
316 (Carolina Biologicals, Burlington, NC) supplemented with banana. Propionic acid and a pupation
317 substrate (Kimwipes Delicate Tasks, Irving TX) were added to each vial. Females were allowed
318 to produce progeny and imported using USDA permit P526P-15-02964. The identity of the
319 species was established by looking at the taxonomical traits of the males produced from

isofemale lines (genital arches, number of sex combs) and the female mating choice (i.e., whether they chose *D. simulans or D. sechellia* in two-male mating trials).

Sequence data and variant calling and phasing

We obtained sequence data from 20 *D. simulans* inbred lines (Rogers et al. 2014) from NCBI's Short Read Archive (BioProject number PRJNA215932). We also sequenced wild-caught outbred *D. sechellia* individuals (see above) from Praslin (*n*=7 diploid genomes), La Digue (*n*=7), Marianne (*n*=2), and Mahé (*n*=7). These new *D. sechellia* genomes are available on the Short Read Archive (BioProject number PRJNA395473). For each line we then mapped all reads with bwa 0.7.15 using the BWA-MEM algorithm (Li 2013) to the March 2012 release of the *D. simulans* assembly produced by Hu et al. (2013) and also used the accompanying annotation based on mapped FlyBase release 5.33 gene models (Gramates et al. 2017). Next, we removed duplicate fragments using Picard (https://github.com/broadinstitute/picard), before using GATK's (version 3.7; McKenna et al. 2010; DePristo et al. 2011; Auwera et al. 2013) HaplotypeCaller in discovery mode with a minimum Phred-scaled variant call quality threshold (-stand_call_conf) of 30. We then used this set of high-quality variants to perform base quality recalibration (with GATK's BaseRecalibrator program), before again using the HaplotypeCaller in discovery mode on the recalibrated alignments. For this second iteration of variant calling we used the --emitRefConfidence GVCF option in order to obtain confidence scores for each site in the genome, whether polymorphic or invariant. Finally, we used GATK's GenotypeGVCFs program to synthesize variant calls and confidences across all individuals and produce genotype calls for each site by setting the --includeNonVariantSites flag, before inferring the most probable haplotypic phase using SHAPEIT v2.r837 (Delaneau et al. 2013). The genotyping and phasing steps were performed separately for our *D. simulans* and *D. sechellia* data, and for each of step in the pipeline outlined above we used default parameters unless otherwise noted. In order to remove potentially erroneous genotypes (at either polymorphic or invariant sites), we considered genotypes as missing data if they had a quality score lower than 20, or were heterozygous in *D. simulans*. After throwing out low-confidence genotypes, we masked all sites in the genome missing genotypes for more than 10% of individuals in either species' population sample, as well as those lying within repetitive elements as predicted by RepeatMasker (http://www.repeatmasker.org). Only SNP calls were included in our downstream analyses (i.e. indels of any size were ignored).

Demographic inference

Having obtained genotype data for our two population samples, we used ∂a∂i to model their shared demographic history on the basis of the folded joint site frequency spectrum (downsampled to *n*=18 and *n*=12 in *D. simulans* and *D. sechellia*, respectively); using the folded spectrum allowed us to circumvent the step of producing whole genome alignments to outgroup

species in *D. simulans* coordinate space in order to attempt to infer ancestral states. We used an isolation-with-migration (IM) model that allowed for continual exponential population size change in each daughter population following the split. This model includes parameters for the ancestral population size (N_{anc}), the initial and final population sizes for *D. simulans* (N_{sim_0} and N_{sim}, respectively), the initial and final sizes for *D. sechellia* (N_{sech_0} and N_{sech}, respectively), the time of the population split (T_D), the rate of migration from *D. simulans* to *D. sechellia* ($m_{sim \rightarrow sech}$), and the rate of migration from *D. sechellia* to *D. simulans* ($m_{sech \rightarrow sim}$). We also fit our data to a pure isolation model that was identical to our IM model but with $m_{sim \rightarrow sech}$ and $m_{sech \rightarrow sim}$ fixed at zero. Our optimization procedure consisted of an initial optimization step using the Augmented Lagrangian Particle Swarm Optimizer (Jansen and Perez 2011), followed by a second step of optimization refining the initial model using the Sequential Least Squares Programming algorithm (Kraft 1988), both of which are included in the pyOpt package for optimization in Python (version 1.2.0; Perez et al. 2012) as in Schrider et al. (2016). We performed ten optimization runs fitting both of these models to our data, each starting from a random initial parameterization, and assessed the fit of each optimization run using the AIC score. Code for performing these optimizations can be obtained from https://github.com/kern-lab/miscDadiScripts, wherein 2popIM.py and 2popIsolation.py fit the IM and isolation models described above, respectively. For scaling times by years rather than numbers of generations, we assumed a generation time of 15 gen/year as has been estimated in *D. melanogaster* (Pool 2015).

Training FILET to detect introgression between *D. simulans* and *D. sechellia*

Having obtained a demographic model that provided an adequate fit to our data, we set out to simulate training examples under this demographic history for each of our three classes (i.e. no migration, migration from *D. simulans* to *D. sechellia*, and from *D. sechellia* to *D. simulans*). For training examples including introgression, T_M was drawn uniformly from the range between zero generations ago and $T_D/4$, while P_M raged uniformly from (0, 1.0]. In addition, in order to make our classifier robust to uncertainty in other parameters in our model, for each training example we drew values of each of the remaining parameters from $[x-(x/2), x+(x/2)]$, where x is our point estimate of the parameter from ∂a∂i. In addition to the parameters from our demographic model (T_D, ρ, N_{anc}, N_{sim}, and N_{sech}), these include the population mutation rate $\theta=4N\mu$ (where μ was set to 3.5×10^{-9}), and the ratio of θ to the population recombination rate ρ (which we set to 0.2). Continuous migration rates were set to zero (i.e. the only migration events that occurred were those governed by the T_M and P_M parameters, and the no-migration examples were truly free of migrants). In total, this training set comprised of 10^4 examples from each of our three classes.

As described above, we masked genomic positions having too many low confidence genotypes or lying within repetitive elements (described above) before proceeding with our classification pipeline. While doing so, we recorded which sites were masked within each 10 kb window in the genome that we would later attempt to classify. In order to ensure that our masking procedure affected our simulated training data in the same manner as our real data, for

each simulated 10 kb window we randomly selected a corresponding window from our real dataset and masked the same sites in the simulated window that had been masked in the real one. We then trained our classifier in the same manner as described above.

In order to ensure that this classifier would indeed be able to reliably uncover loci experiencing recent gene flow between our two populations, we assessed its performance on simulated test data. First, we applied the classifier to test examples simulated under this same model (again, 10^4 for each class). Next, to address the effect of demographic model misspecification, we applied our classifier to an isolation model with a different parameterization and no continuous size change in the daughter populations. For this model we simply set N_{sim} and N_{sech} to $\pi_{sim}/4\mu$ and $\pi_{sech}/4\mu$, respectively, where π for a species is the average nucleotide diversity among all windows included in our analysis after filtering, and μ was again set to 3.5×10^{-9}. We then set N_{anc} to be equal to N_{sim}, and set T to $d_{xy}/(2\mu) - 2N_{anc}$ generations where d_{xy} is the average divergence between *D. simulans* and *D. sechellia* sequences across all windows. This latter value is simply the expected TMRCA for cross-species pairs of genomes, minus the expected waiting time until coalescence during the one-population (i.e. ancestral) phase of the model. This simple model thus produces samples with similar levels of nucleotide diversity for the two daughter populations as those produced under our IM model, but that would differ in other respects (e.g. the joint site frequency spectrum and linkage disequilibrium, which would be affected by continuous population size change after the split). For both test sets we masked sites in the same manner as for our training data before running FILET.

Classifying genomic windows with FILET

We examined 10 kb windows in the *D. simulans* and *D. sechellia* genomes, summarizing diversity in the joint sample with the same feature vector as used for classification (see above), ignoring sites that were masked as described above. We omitted from this analysis any window for which >25% of sites were masked, and then applied our classifier to each remaining window, calculating posterior class membership probabilities for each class. We then used a simple clustering algorithm to combine adjacent windows showing evidence of introgression into contiguous introgressed elements: we obtained all stretches of consecutive windows with >90% probability of introgression as predicted by FILET (i.e. the probability of no-introgression class <10%), and retained as putatively introgressed regions those that contained at least one window with >95% probability of introgression. In order to test for enrichment of these introgressed regions for genic/intergenic sequence or particular Gene Ontology (GO) terms from the FlyBase 5.33 annotation release (Gramates et al. 2017), we performed a permutation test in which we randomly assigned a new location for each cluster or introgressed windows (ensuring the entire permuted cluster landed within accessible windows of the genome according to our data filtering criteria). We generated 10,000 of these permutations.

RESULTS AND DISCUSSION

440
441 **FILET detects introgressed loci with high sensitivity and specificity**
442
443 We sought to devise a bioinformatic approach capable of leveraging population genomic data
444 from two related population samples to uncover introgressed loci with high sensitivity and
445 specificity. In the Materials and Methods, we describe several previous and novel statistics
446 designed to this end. However, rather than preoccupying ourselves with the search for the ideal
447 statistic for this task, we took the alternative approach of assembling a classifier leveraging many
448 statistics that would in concert have greater power to discriminate between introgressed and non-
449 introgressed loci. Supervised machine learning methods have proved highly effective at making
450 inferences in high-dimensional datasets. In this vein, we designed FILET, which uses an
451 extension of random forests called an Extra-Trees classifier (Geurts et al. 2006). We previously
452 succeeded in applying Extra-Trees classifiers for a separate population genetic task—finding
453 recent positive selection and discriminating between hard and soft sweeps (Schrider and Kern
454 2016; Schrider and Kern 2017)—though other methods such as support vector machines (Cortes
455 and Vapnik 1995) or deep learning (LeCun et al. 2015) could also be applied to this task.
456 Briefly, FILET assigns a given genomic window to one of three distinct classes—recent
457 introgression from population 1 into population 2, introgression from population 2 into 1, or the
458 absence of introgression—on the basis of a vector of summary statistics that contain information
459 about the two-population sample's history. This feature vector contains a variety of statistics
460 summarizing patterns of diversity within each population sample, as well as a number of
461 statistics examining cross-population variation (see Materials and Methods for a full description).
462 FILET must be trained to distinguish among these three classes, which we accomplish by
463 supplying 10,000 simulated example genomic windows of each class, with each example
464 represented by its feature vector. Because we expect that the majority of introgression events to
465 be non-adaptive, these simulations did not include natural selection. Once this training is
466 complete, FILET can then be used to infer the class membership of additional genomic windows,
467 whether from simulated or real data.
468 We began by assessing FILET's power on a number of simulated datasets, examining
469 windows roughly equivalent to 10 kb in length in *Drosophila* (Materials and Methods). In
470 particular, because the power to detect introgression depends on the time since their divergence,
471 T_D, we measured FILET's performance under four different values of T_D, training a separate
472 classifier for each. In Figure 1 (T_D=0.25×4N) and Figure S3 (T_D values of 1, 4, and 16×4N), we
473 compare FILET's power to that of two related statistics that have been devised to detect
474 introgressed windows, d_{min} and G_{min} (Materials and Methods). These figures show that FILET
475 has high sensitivity to introgression across a much wider range of introgression timings (T_M) and
476 intensities (P_M) than either of these statistics under each value of T_D, and that this disparity is
477 amplified dramatically for smaller values of T_D. Furthermore, these figures demonstrate that
478 FILET infers the correct directionality of recent introgression with high accuracy, whereas d_{min}
479 and G_{min} contain no information about the direction of gene flow. Finally, FILET does not appear

480 especially sensitive when the source of gene flow is an unsampled ghost population rather than
481 one of the two sequenced populations (Figure S4), though it could potentially be trained to detect
482 such cases if desired.

483 We also note that for d_{min} and G_{min} we established 95% significance thresholds from our
484 simulated training data without introgression, thereby achieving a false positive rate of 5%. In
485 order to assess FILET's false positive rate, we classified a set of test simulations without
486 introgression and found that FILET's false positive rate was considerably lower (Figure 1 and
487 Figure S3) except for our largest value of T_D, where it was comparable (0.4% for T_D=0.25×4N
488 but ~6% for T_D=16×4N). Thus, FILET achieves much greater sensitivity to introgression than
489 d_{min} and G_{min} often at a much lower false positive rate. We also demonstrate the FILET's greater
490 power than these statistics via ROC curves (Figure S5), where it outperforms each statistic under
491 each scenario. Specifically, the difference in power between FILET and d_{min} is dramatic for
492 smaller values of T_D (area under curve, or AUC, of 0.85 versus 0.73 when T_D=0.25×4N for
493 FILET and d_{min}, respectively) but comparatively miniscule for our largest T_D (AUC of 0.94
494 versus 0.93 when T_D=16×4N). It therefore appears that FILET's performance gain relative to
495 single statistics is highest for the more difficult task of finding introgression between very
496 recently diverged populations, while for the easier case of detecting introgression between highly
497 diverged populations some single statistics may perform nearly as well.

498 We also experimented with smaller training sets, finding similar classification power
499 (measured by AUC) as above when we trained FILET using only 1000 or even 100 simulated
500 examples per class (Figure S6), though in the latter case estimated class posterior probabilities
501 were far less accurate. In addition, we examined the effect of altering the target window size
502 used when training FILET (Figure S7). As expected, there is a tradeoff between accuracy and
503 resolution, though 5 or 10 kb windows appear balance this tradeoff fairly well.

504 While FILET is designed for identifying particular genomic windows that experienced
505 introgression as a result of a pulse migration event, genomic windows with genealogies that
506 include introgression may of course also be produced by continuous migration, with the timing
507 of geneflow varying from window to window. We therefore simulated genomic windows
508 experiencing a variety of bidirectional migration rates under each of our values of T_D and
509 recorded the fraction of windows for which our sampled individuals contained at least one
510 migrant lineage. Next, for each simulated window we applied the FILET classifier trained under
511 the appropriate divergence time as described above, recording the fraction of windows with at
512 least one migrant lineage that were classified as introgressed by FILET. The results of these tests
513 (Table S1) show that for each value of T_D, the lowest bidirectional migration rates that we tested
514 do not produce migrant lineages, while higher rates will produce a small to modest fraction
515 migrants, most of which are undetected (e.g. when m=0.01, 23% and 59% of windows contain at
516 least one migrant when T_D=0.25 and T_D=1, respectively, but <5% are detected by FILET). Thus,
517 FILET, as currently trained, may not be sensitive to gene flow produced by low levels of
518 continuous migration. However, as the migration rate increases further, more and more of these

519 migrant windows will be detected (e.g. when m=1, 70% and 100% of windows are detected as
520 migrants by FILET when T_D=0.25 and T_D=1, respectively).
521 Although our goal was to use a set of statistics to perform more accurate inference than
522 possible using individual ones, another benefit of our Extra-Trees approach is that it also allows
523 for a natural way to evaluate the extent to which different statistics are informative under
524 different scenarios of introgression. To this end, we used the Extra-Trees classifier to calculate
525 feature importance, which captures the ability of each statistic to separate the data into its
526 respective classes (Materials and Methods). We find that for our lowest T_D (a split N generations
527 ago) the top four features, all with similar importance, are the density of private alleles in
528 population 1, the density of private alleles in population 2, $d_{d\text{-}Rank1}$, and $d_{d\text{-}Rank2}$. For our next-
529 lowest T_D (4N generations ago), the top four, again with similar importance score estimates, are
530 F_{ST}, Z_X, d_{d1}, and d_{d2}. Thus our newly devised d_d and Z_X statistics seem to be particularly
531 informative in the case of recent introgression between closely related populations. For the larger
532 values of T_D, d_{xy} and d_{min} rise to prominence. The complete lists of feature importance for each
533 T_D are shown in Table S2.
534 The exceptional accuracy with which FILET uncovers introgressed loci underscores the
535 potential for machine learning methods to yield more powerful population genetic inferences
536 than can be achieved via more conventional tools which are often based on a single statistic.
537 Indeed, machine learning tools have been successfully leveraged in efforts to detect recent
538 positive selection (Pavlidis et al. 2010; Lin et al. 2011; Ronen et al. 2013; Pybus et al. 2015;
539 Schrider and Kern 2016), to infer demographic histories (Pudlo et al. 2016), or even to perform
540 both of these tasks concurrently (Sheehan and Song 2016).
541
542 **Joint demographic history of *D. simulans* and *D. sechellia***
543
544 As described in the Materials and Methods, we used publically available *D. simulans* sequence
545 data (Rogers et al. 2014), and collected and sequenced a set of *D. sechellia* genomes. We
546 mapped reads from these genomes to the *D. simulans* assembly (Hu et al. 2013), obtaining high
547 coverage >28× for each sequence (see sampling locations, mapping statistics, and Short Read
548 Archive identifier information listed in Table S3). We do not expect that our reliance on the *D.*
549 *simulans* assembly resulted in any appreciable bias, as reads from our *D. sechellia* genomes were
550 successfully mapped to the reference genome at nearly the same rate as reads from *D. simulans*
551 (Table S3).
552 After completing variant calling and phasing (Materials and Methods), we performed
553 principal components analysis on intergenic SNPs at least 5 kb away from the nearest gene in
554 order to mitigate the bias introduced by linked selection (Gazave et al. 2014; Schrider et al.
555 2016), and observed evidence of population structure within *D. sechellia*. In particular, the
556 samples obtained from Praslin clustered together, while all remaining samples formed a separate
557 cluster (Figure S8A). Running fastStructure (Raj et al. 2014) on this same set of SNPs yielded
558 similar results: when the number of subpopulations, K, was set to 2 (the optimal value for K

559 selected by fastStructure's chooseK.py script), our data were again subdivided into Praslin and
560 non-Praslin clusters. Indeed, across all values of *K* between 2 and 8, fastStructure's results were
561 suggestive of marked subdivision between Praslin and non-Praslin samples, and comparatively
562 little subdivision within the non-Praslin data (Figure S8B). This surprising result differs
563 qualitatively from previous observations from smaller numbers of loci (Legrand et al. 2009;
564 Legrand et al. 2011), and underscores the importance of using data from many loci—preferably
565 intergenic and genome-wide—in order to infer the presence or absence of population structure.
566 Next, we examined the site frequency spectra of the Praslin and non-Praslin clusters,
567 noting that both had an excess of intermediate frequency alleles in comparison to that of the *D.*
568 *simulans* dataset (Figure S9), in line with our expectations of *D. sechellia's* demographic history.
569 We also note that the Praslin samples contained far more variation (50,243 sites were
570 polymorphic within Praslin) than non-Praslin samples (4,108 SNPs within these samples). This
571 difference in levels of variation may reflect a much lesser degree of population structure and/or
572 inbreeding on the island of Praslin than across the other islands, or may result from other
573 demographic processes. Additional samples from across the Seychelles would be required to
574 address this question. In any case, in light of this observation we limited our downstream
575 analyses of *D. sechellia* sequences to those from Praslin.
576 Because we required a model from which to simulate training data for FILET, we next
577 inferred a joint demographic history of our population samples using ∂a∂i (Gutenkunst et al.
578 2009). In particular, we fit two demographic models to our dataset: an isolation-with-migration
579 (IM) model allowing for continuous population size change and migration following the
580 population divergence, and an isolation model with the same parameters but fixing migration
581 rates at zero (Materials and Methods). In Table S4 we show our model optimization results,
582 which show clear support for the IM model over the isolation model. The IM model that
583 provided the best fit to our data (Figure 2A) includes a much larger population size in *D*
584 *simulans* than *D. sechellia* (a final size of 9.3×10^6 for *D. simulans* versus 2.6×10^4 for *D.*
585 *sechellia*), consistent with the much greater diversity levels in *D. simulans* (Begun et al. 2007;
586 Legrand et al. 2009). This model also exhibits a modest rate of migration, with a substantially
587 higher rate of gene flow from *D. simulans* to *D. sechellia* ($2\times N_{anc}m$=0.086) than vice-versa
588 ($2\times N_{anc}m$=0.013). Thus, the results of our demographic modeling are consistent with the
589 observation of hybrid males in the Seychelles (Matute and Ayroles 2014), and the possibility of
590 recent introgression between these two species across a substantial fraction of the genome (see
591 Garrigan et al. 2012; Navascués et al. 2014).
592 An interesting characteristic of the model shown in Figure 2A is that, assuming 15
593 generations per year, the estimated time of the *D. simulans*- *D. sechellia* population split is ~86
594 kya, or 1.3×10^6 generations ago, in stark contrast to a recent estimate of the of 2.5×10^6
595 generations ago from Garrigan et al. (2012) which was not based on population genomic data,
596 but rather on single genomes. Supporting our inference, we note that our average intergenic
597 cross-species divergence of 0.017 yields an average TMRCA of ~2.5×10^6 generations ago,
598 assuming a mutation rate of 3.5×10^{-9} mutations per generation as observed in *D. melanogaster*

599 (Keightley et al. 2009; Schrider et al. 2013), and this estimate would include the time before
600 coalescence in the ancestral population. Unless the mutation rate the *D. simulans* species
601 complex is substantially lower than in *D. melanogaster*, a population split time of 2.5×10^6
602 generations ago therefore seems unlikely given that the ancestral population size (and therefore
603 the period of time between the population divergence and average TMRCA) was probably large
604 (>500,000 by our estimate). Thus, we conclude that the *D. simulans* and *D. sechellia* populations
605 may have diverged more recently than previously appreciated, perhaps within the last 100,000
606 years.
607 Although the specific parameterization of our model should be regarded as a preliminary
608 view of these species' demographic history that is adequate for the purposes of training FILET,
609 future efforts with larger sample sizes will be required to refine this model. That being said, the
610 basic features of this model—a much larger *D. simulans* population size than *sechellia*, and a
611 fairly large ancestral population size—are unlikely to change qualitatively.
612
613 **Widespread introgression from *D. simulans* to *D. sechellia***
614
615 *Accuracy and robustness of FILET under estimated model:* Having obtained a suitable model of
616 the *D. simulans*- *D. sechellia* joint demographic history, we proceeded to simulate training data
617 and train FILET for application to our dataset (Materials and Methods). After training FILET
618 and applying it to simulated data under the estimated demographic model, we find that we have
619 good sensitivity to introgression (56% of windows with introgression are detected, on average),
620 and a false positive rate of only 0.2% (Figure 2B). Thus, while we may miss some introgressed
621 loci, we can have a great deal of confidence in the events that we do recover. Our feature
622 rankings for this classifier are included in Table S2—under this scenario the most informative
623 feature is our newly devised $d_{d\text{-}sim}$. Note that we achieve high accuracy despite some of the
624 difficulties presented by the demographic model in Figure 2A, most notably the asymmetry in
625 effective population sizes between our two species. Indeed, because our method is trained under
626 this demographic history, the characteristics of genealogies demographic model (such as
627 asymmetry in π) with and without introgression become the signal used by FILET to make its
628 classifications.
629 As shown in Figure 2B we find that this classifier has greater sensitivity to introgression
630 from *D. sechellia* to *D. simulans* than vice-versa. The cause of a stronger signal of *D. sechellia*→
631 *D. simulans* introgression can be understood from a consideration of the d_{min} statistic under each
632 of our three classes. When there is no introgression, d_{min} will be similar to the expected
633 divergence between *D. simulans* and *D. sechellia*; when there is introgression from *D. simulans*
634 to *D. sechellia*, we may expect d_{min} to be proportional to π_{sim}, which may only be a moderate
635 reduction relative to the no-introgression case given the large population size in *D. simulans*;
636 when there is introgression from *D. sechellia* to *D. simulans* then d_{min} is proportional to π_{sech}
637 which is dramatically lower than the expectation without introgression. While many of our
638 statistics do not rely on d_{min}, this example illustrates an important property of the genealogy of

639 introgression from *D. sechellia* to *D. simulans* that would make it easier to detect than gene flow
640 in the reverse direction.
641 We also tested this classifier's performance on a different demographic scenario (Table
642 S4) in order to examine the effect of model misspecification during training. In particular, we
643 devised a simple island model with two population sizes: a larger size for *D. simulans* and the
644 ancestral population (7.6×10^5), and a smaller size for *D. sechellia* (5.7×10^4) with a split time of
645 ~59 kya. Our simple procedure for estimating these values is described in the Materials and
646 Methods. Again, we find that we have good power to detect introgression with a very low false
647 positive rate (0.28%; Figure S10). Although there are myriad incorrect models that we could test
648 FILET against, this example suggests that FILET is robust to demographic misspecification.
649
650 *Application to population genomic data:* We applied FILET to 10,185 non-overlapping 10 kb
651 windows that passed our data quality filters (101.85 Mb in total, or 86.7% of the five major
652 chromosome arms; Materials and Methods). FILET classified 267 windows as introgressed with
653 high-confidence, which we clustered into 94 contiguous regions accounting for 2.93% of the
654 accessible portion of the genome (2.99 Mb in total; Materials and Methods). This finding is
655 qualitatively similar to a previous estimate (4.6%) by Garrigan et al. (2012) based on
656 comparisons of single genomes from each species in the *D. simulans* complex. Unlike this
657 previous effort, FILET is able to infer the directionality of introgression with high confidence
658 (Figure 2B), and we find evidence that the majority of this introgression has been in the direction
659 of *D. simulans* to *D. sechellia*: only 21 of the 267 (7.9%) putatively introgressed windows were
660 classified as introgressed from *sechellia* to *D. simulans*. This finding is not a result of a detection
661 bias, as we have greater power to detect gene flow from *D. sechellia* to *D. simulans* than in the
662 reverse direction. Given that our *D. simulans* sequences are from the mainland, one interpretation
663 of this result is that although there has been recent gene flow from *D. simulans* into the
664 Seychelles, where *D. simulans* and *D. sechellia* occasionally hybridize, there does not appear to
665 be an appreciable rate of back-migration to the mainland of *D. simulans* individuals harboring
666 haplotypes donated from *D. sechellia*. On the other hand, *D. sechellia* alleles may often be
667 purged from *D. simulans* by natural selection. This may be in part due to the reduced ecological
668 niche size of *D. sechellia*, such that any alleles which may introgress into *D. simulans* and lead to
669 preference for or resistance to *Morinda* fruit may prove deleterious in other environments. More
670 generally, *D. sechellia* haplotypes introgressing into *D. simulans* may harbor more deleterious
671 alleles due to their smaller population size, which will be more effectively purged in the larger *D.*
672 *simulans* population if mutations are not fully recessive (Harris and Nielsen 2016). Tests of these
673 hypotheses will have to wait for a population sample of genomes from *D. simulans* collected in
674 the Seychelles.
675 We asked whether our candidate introgressed loci were enriched for particular GO terms
676 using a permutation test (Materials and Methods), finding no such enrichment. We did observe a
677 significant deficit in the number of genes either partially overlapping or contained entirely within
678 introgressed regions in our true set versus the permuted set (297 vs. 373.2, respectively;

679 *P*=0.083; one-sided permutation test). This paucity of introgressed genes is consistent with
680 introgressed functional sequence often being deleterious.
681 One notable introgressed region on 3R that FILET identified had been previously found
682 by Garrigan et al. as containing a 15 kb region of introgression. We show that gene flow in this
683 region actually extends for over 200 kb (Figure 3). When Brand et al. (2013) sequenced the 15
684 kb region originally flagged by Garrigan et al. in a number of *D. simulans* and *D. sechellia*
685 individuals, they also uncovered evidence of a selective sweep in *D. sechellia* originating from
686 an adaptive introgression from *D. simulans*. Our data set also supports the presence of an
687 adaptive introgression event at this locus: a 10 kb window lying within the putative sweep region
688 (highlighted in Figure 3) is in the lower 5% tail of both d_{min} (consistent with introgression) and
689 π_{sech} (consistent with a sweep in *sechellia*); this is the only window in the genome that is in the
690 lower 5% tail for both of these statistics. This region contains two ionotropic glutamate
691 receptors, *CG3822* and *Ir93a*, which may be involved in chemosensing among other functions
692 (Benton et al. 2009), and the latter of which appears to play a role in resistance to
693 entomopathogenic fungi (Lu et al. 2015). Also near the trough of variation within *D. sechellia*
694 are several members of the *Turandot* gene family, which are involved in humoral stress
695 responses to various stressors including heat, UV light, and bacterial infection (Ekengren and
696 Hultmark 2001; Ekengren et al. 2001), and perhaps parasitoid attack as well (Salazar-Jaramillo et
697 al. 2017). On the other hand, Brand et al. (2013) hypothesize that the target of selection may be a
698 transcription factor binding hotspot between *RpS30* and *CG15696*, and the phenotypic target of
699 this sweep remains unclear.
700 Interestingly, this particular window is the only one in this region that is classified by
701 FILET as having recent gene flow from *D. sechellia* to *D. simulans*. However this classification
702 may be erroneous as one might expect FILET, which was not trained on any examples of
703 adaptive introgression, to make an error in such a scenario because rather than gene flow
704 increasing polymorphism in the recipient population, diversity is greatly diminished if the
705 introgressed alleles rapidly sweep toward fixation. We note that this window is immediately
706 flanked by a large number of windows classified as introgressed from *D. simulans* to *D. sechellia*
707 and which show a large increase in diversity in the recipient population as expected. Moreover,
708 Brand et al.'s phylogenetic analysis of introgression in this region also supported gene flow in
709 this direction. Brand et al. also found evidence suggesting that the introgressed haplotype began
710 sweeping to higher frequency in *D. simulans* (though it has not reached fixation in this species)
711 prior to the timing of the introgression and subsequent sweep in *D. sechellia*. Thus we conclude
712 that the adaptive allele probably did indeed originate in *D. simulans* before migrating to *D.*
713 *sechellia*, and FILET's apparent error in this case underscores the genealogical differences
714 between adaptive gene flow and introgression events involving only neutral alleles.
715
716 **Concluding remarks**
717

Here we present a novel machine learning approach, FILET, that leverages population genomic data from two related populations in order to determine whether a given genomic window has experienced gene flow between these populations, and if so in which direction. We applied FILET to a set of *D. simulans* genomes as well as a new set of whole genome sequences from the closely related island endemic *D. sechellia*, confirming widespread introgression and also inferring that this introgression was largely in the direction of *D. simulans* to *D. sechellia*. Future work leveraging *D. simulans* data sampled from the Seychelles will be required to determine whether this asymmetry is a consequence of low rate of migration of *D. simulans* back to mainland Africa (where our *D. simulans* data were obtained), or whether the directionality of gene flow is biased on the islands themselves. In addition to creating FILET, we devised several new statistics, including the d_d statistics and Z_X which our feature rankings show to be quite useful for uncovering gene flow. Despite the success of FILET on both simulated data sets and real data from *Drosophila*, there are several improvements that could be made. First, by framing the problem as one of parameter estimation (i.e. regression) rather than classification, we may be able to precisely infer the values of relevant parameters of introgression events (i.e. the time of the event and the number of migrant lineages). Deep learning methods, which naturally allow for both classification and regression, may prove particularly useful for this task (LeCun et al. 2015). Indeed, Sheehan and Song (2016) used deep learning to infer demographic parameters (regression) while simultaneously identifying selective sweeps (classification). Another step we have not taken is to explicitly handle adaptive introgression, which could potentially greatly improve our approach's power to detect such events.

While population genetic inference has traditionally relied on the design of a summary statistic sensitive to the evolutionary force of interest, a number of highly successful supervised machine learning methods have been put forth within the last few years (Pavlidis et al. 2010; Lin et al. 2011; Ronen et al. 2013; Pybus et al. 2015; Pudlo et al. 2016; Schrider and Kern 2016; Sheehan and Song 2016). As genomic data sets continue to grow, we argue that machine learning approaches leveraging high dimensional feature spaces have the potential to revolutionize evolutionary genomic inference.

ACKNOWLEDGMENTS

We thank Michael Lan for his work on an early iteration of this project. D.R.S. was supported by NIH award no. K99HG008696. A.D.K. was supported in part by NIH award no. R01GM078204. D.R.M was supported by R01GM121750.

REFERENCES

Andrade López J, Lanno S, Auerbach J, Moskowitz E, Sligar L, Wittkopp P and Coolon J. 2017. Genetic basis of octanoic acid resistance in Drosophila sechellia: functional analysis of a fine-mapped region. *Mol Ecol* 26: 1148-1160.

758 Auwera GA, Carneiro MO, Hartl C, et al. 2013. From FastQ data to high-confidence variant
759 calls: the genome analysis toolkit best practices pipeline. *Current protocols in*
760 *bioinformatics* 43: 11.10. 11-11.10. 33.
761 Barton NH and Hewitt GM. 1985. Analysis of hybrid zones. *Annual review of Ecology and*
762 *Systematics* 16: 113-148.
763 Begun DJ, Holloway AK, Stevens K, et al. 2007. Population genomics: whole-genome analysis
764 of polymorphism and divergence in Drosophila simulans. *PLoS Biol* 5: e310.
765 Benton R, Vannice KS, Gomez-Diaz C and Vosshall LB. 2009. Variant ionotropic glutamate
766 receptors as chemosensory receptors in Drosophila. *Cell* 136: 149-162.
767 Brand CL, Kingan SB, Wu L and Garrigan D. 2013. A selective sweep across species boundaries
768 in Drosophila. *Mol Biol Evol* 30: 2177-2186.
769 Brandvain Y, Kenney AM, Flagel L, Coop G and Sweigart AL. 2014. Speciation and
770 introgression between Mimulus nasutus and Mimulus guttatus. *PLoS Genet* 10:
771 e1004410.
772 Breiman L. 2001. Random forests. *Machine Learning* 45: 5-32.
773 Breiman L, Friedman J, Stone CJ and Olshen RA. 1984. Classification and regression trees: CRC
774 press.
775 Chan AH, Jenkins PA and Song YS. 2012. Genome-wide fine-scale recombination rate variation
776 in Drosophila melanogaster. *PLoS Genet* 8: e1003090.
777 Cortes C and Vapnik V. 1995. Support-vector networks. *Machine Learning* 20: 273-297.
778 Dekker T, Ibba I, Siju K, Stensmyr MC and Hansson BS. 2006. Olfactory shifts parallel
779 superspecialism for toxic fruit in Drosophila melanogaster sibling, D. sechellia. *Curr Biol*
780 16: 101-109.
781 Delaneau O, Zagury J-F and Marchini J. 2013. Improved whole-chromosome phasing for disease
782 and population genetic studies. *Nat Methods* 10: 5-6.
783 DePristo MA, Banks E, Poplin R, et al. 2011. A framework for variation discovery and
784 genotyping using next-generation DNA sequencing data. *Nat Genet* 43: 491-498.
785 Ekengren S and Hultmark D. 2001. A family of Turandot-related genes in the humoral stress
786 response of Drosophila. *Biochem Biophys Res Commun* 284: 998-1003.
787 Ekengren S, Tryselius Y, Dushay MS, Liu G, Steiner H and Hultmark D. 2001. A humoral stress
788 response in Drosophila. *Curr Biol* 11: 714-718.
789 Farine J-P, Legal L, Moreteau B and Le Quere J-L. 1996. Volatile components of ripe fruits of
790 Morinda citrifolia and their effects on Drosophila. *Phytochemistry* 41: 433-438.
791 Fay JC and Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405-
792 1413.
793 Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, Dambroski H, Filchak KE and Aluja M.
794 2005. Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in
795 Rhagoletis. *Proceedings of the National Academy of Sciences* 102: 6573-6580.
796 Fontaine MC, Pease JB, Steele A, et al. 2015. Extensive introgression in a malaria vector species
797 complex revealed by phylogenomics. *Science* 347: 1258524.
798 Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR and Presgraves DC.
799 2012. Genome sequencing reveals complex speciation in the Drosophila simulans clade.
800 *Genome Res* 22: 1499-1511.
801 Gazave E, Ma L, Chang D, et al. 2014. Neutral genomic regions refine models of recent rapid
802 human population growth. *Proceedings of the National Academy of Sciences* 111: 757-
803 762.

804 Geneva AJ, Muirhead CA, Kingan SB and Garrigan D. 2015. A new method to scan genomes for
805 introgression in a secondary contact model. *PLoS ONE* 10: e0118621.
806 Geurts P, Ernst D and Wehenkel L. 2006. Extremely randomized trees. *Machine Learning* 63: 3-
807 42.
808 Gramates LS, Marygold SJ, Santos Gd, et al. 2017. FlyBase at 25: looking to the future. *Nucleic*
809 *Acids Res* 45: D663-D671.
810 Green RE, Krause J, Briggs AW, et al. 2010. A draft sequence of the Neandertal genome.
811 *Science* 328: 710-722.
812 Gutenkunst RN, Hernandez RD, Williamson SH and Bustamante CD. 2009. Inferring the joint
813 demographic history of multiple populations from multidimensional SNP frequency data.
814 *PLoS Genet* 5: e1000695.
815 Harris K and Nielsen R. 2016. The genetic cost of Neanderthal introgression. *Genetics* 203: 881-
816 891.
817 Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new
818 mutation and standing variation as sources of adaptive variation. *Mol Ecol* 22: 4606-
819 4618.
820 Hey J and Kliman RM. 1993. Population genetics and phylogenetics of DNA sequence variation
821 at multiple loci within the Drosophila melanogaster species complex. *Mol Biol Evol* 10:
822 804-822.
823 Hu TT, Eisen MB, Thornton KR and Andolfatto P. 2013. A second-generation assembly of the
824 *Drosophila simulans* genome provides new insights into patterns of lineage-specific
825 divergence. *Genome Res* 23: 89-98.
826 Huang Y and Erezyilmaz D. 2015. The genetics of resistance to Morinda fruit toxin during the
827 postembryonic stages in Drosophila sechellia. *G3: Genes, Genomes, Genetics* 5: 1973-
828 1981.
829 Hudson RR. 2000. A new statistic for detecting genetic differentiation. *Genetics* 155: 2011-2014.
830 Hudson RR, Slatkin M and Maddison W. 1992. Estimation of levels of gene flow from DNA
831 sequence data. *Genetics* 132: 583-589.
832 Huerta-Sánchez E, Jin X, Bianba Z, et al. 2014. Altitude adaptation in Tibetans caused by
833 introgression of Denisovan-like DNA. *Nature* 512: 194-197.
834 Hungate EA, Earley EJ, Boussy IA, Turissini DA, Ting C-T, Moran JR, Wu M-L, Wu C-I and
835 Jones CD. 2013. A locus in Drosophila sechellia affecting tolerance of a host plant toxin.
836 *Genetics* 195: 1063-1075.
837 Jansen PW and Perez RE. 2011. Constrained structural design optimization via a parallel
838 augmented Lagrangian particle swarm optimization approach. *Computers & Structures*
839 89: 1352-1366.
840 Joly S, McLenachan PA and Lockhart PJ. 2009. A statistical approach for distinguishing
841 hybridization and incomplete lineage sorting. *The American Naturalist* 174: E54-E70.
842 Jones CD. 1998. The genetic basis of Drosophila sechellia's resistance to a host plant toxin.
843 *Genetics* 149: 1899-1908.
844 Jones CD. 2005. The genetics of adaptation in Drosophila sechellia. *Genetica* 123: 137.
845 Juric I, Aeschbacher S and Coop G. 2016. The strength of selection against Neanderthal
846 introgression. *PLoS Genet* 12: e1006340.
847 Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S and Blaxter M. 2009. Analysis of the
848 genome sequences of three Drosophila melanogaster spontaneous mutation accumulation
849 lines. *Genome Res* 19: 1195-1201.

850 Kelly JK. 1997. A test of neutrality based on interlocus associations. *Genetics* 146: 1197-1206.
851 Kern AD, Jones CD and Begun DJ. 2004. Molecular population genetics of male accessory
852 gland proteins in the Drosophila simulans complex. *Genetics* 167: 725-735.
853 Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J
854 and Hey J. 2000. The population genetics of the origin and divergence of the Drosophila
855 simulans complex species. *Genetics* 156: 1913-1931.
856 Kraft D. 1988. A software package for sequential quadratic programming: DFVLR
857 Obersfaffeuhofen, Germany.
858 Kulathinal RJ, Stevison LS and Noor MA. 2009. The genomics of speciation in Drosophila:
859 diversity, divergence, and introgression estimated using low-coverage genome
860 sequencing. *PLoS Genet* 5: e1000550.
861 LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. *Nature* 521: 436-444.
862 Legal L, Chappe B and Jallon JM. 1994. Molecular basis ofMorinda citrifolia (L.): Toxicity on
863 drosophila. *J Chem Ecol* 20: 1931-1943.
864 Legal L, Moulin B and Jallon JM. 1999. The relation between structures and toxicity of
865 oxygenated aliphatic compounds homologous to the insecticide octanoic acid and the
866 chemotaxis of two species of Drosophila. *Pestic Biochem Physiol* 65: 90-101.
867 Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D and Cariou M-L. 2009. Species-wide
868 genetic variation and demographic history of Drosophila sechellia, a species lacking
869 population structure. *Genetics* 182: 1197-1206.
870 Legrand D, Vautrin D, Lachaise D and Cariou M-L. 2011. Microsatellite variation suggests a
871 recent fine-scale population structure of Drosophila sechellia, a species endemic of the
872 Seychelles archipelago. *Genetica* 139: 909.
873 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
874 *arXiv*.
875 Lin K, Li H, Schlötterer C and Futschik A. 2011. Distinguishing positive selection from neutral
876 evolution: boosting the performance of summary statistics. *Genetics* 187: 229-244.
877 Louis J and David J. 1986. Ecological specialization in the Drosophila melanogaster species
878 subgroup: a case study of D. sechellia. *Acta oecologica Oecologia generalis* 7: 215-229.
879 Lu H-L, Wang JB, Brown MA, Euerle C and Leger RJS. 2015. Identification of Drosophila
880 mutants affecting defense to an entomopathogenic fungus. *Scientific reports* 5.
881 Mallet J. 2005. Hybridization as an invasion of the genome. *Trends in ecology & evolution* 20:
882 229-237.
883 Martin SH, Dasmahapatra KK, Nadeau NJ, et al. 2013. Genome-wide evidence for speciation
884 with gene flow in Heliconius butterflies. *Genome Res* 23: 1817-1828.
885 Matsuo T, Sugaya S, Yasukawa J, Aigaki T and Fuyama Y. 2007. Odorant-binding proteins
886 OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila
887 sechellia. *PLoS Biol* 5: e118.
888 Matute D and Ayroles J. 2014. Hybridization occurs between Drosophila simulans and D.
889 sechellia in the Seychelles archipelago. *J Evol Biol* 27: 1057-1068.
890 McKenna A, Hanna M, Banks E, et al. 2010. The Genome Analysis Toolkit: a MapReduce
891 framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-
892 1303.
893 Melo MC, Salazar C, Jiggins CD and Linares M. 2009. Assortative mating preferences among
894 hybrids offers a route to hybrid speciation. *Evolution* 63: 1660-1665.

895 Navascués M, Legrand D, Campagne C, Cariou M-L and Depaulis F. 2014. Distinguishing
896 migration from isolation using genes with intragenic recombination: detecting
897 introgression in the Drosophila simulans species complex. *BMC Evol Biol* 14: 89.
898 Neafsey DE, Barker BM, Sharpton TJ, et al. 2010. Population genomic sequencing of
899 Coccidioides fungi reveals recent hybridization and transposon control. *Genome Res* 20:
900 938-946.
901 Nei M and Li W-H. 1979. Mathematical model for studying genetic variation in terms of
902 restriction endonucleases. *Proceedings of the National Academy of Sciences* 76: 5269-
903 5273.
904 Nürnberger B, Lohse K, Fijarczyk A, Szymura JM and Blaxter ML. 2016. Para-allopatry in
905 hybridizing fire-bellied toads (Bombina bombina and B. variegata): Inference from
906 transcriptome-wide coalescence analyses. *Evolution* 70: 1803-1818.
907 Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, McMillan WO
908 and Jiggins CD. 2012. Adaptive introgression across species boundaries in Heliconius
909 butterflies. *PLoS Genet* 8: e1002752.
910 Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T and
911 Reich D. 2012. Ancient admixture in human history. *Genetics* 192: 1065-1093.
912 Pavlidis P, Jensen JD and Stephan W. 2010. Searching for footprints of positive selection in
913 whole-genome SNP data from nonequilibrium populations. *Genetics* 185: 907-922.
914 Pedregosa F, Varoquaux G, Gramfort A, et al. 2011. Scikit-learn: Machine learning in Python.
915 *Journal of Machine Learning Research* 12: 2825-2830.
916 Perez RE, Jansen PW and Martins JR. 2012. pyOpt: a Python-based object-oriented framework
917 for nonlinear constrained optimization. *Structural and Multidisciplinary Optimization* 45:
918 101-118.
919 Pool JE. 2015. The mosaic ancestry of the Drosophila genetic reference panel and the D.
920 melanogaster reference genome reveals a network of epistatic fitness interactions. *Mol*
921 *Biol Evol* 32: 3236-3251.
922 Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M and Robert CP. 2016. Reliable ABC
923 model choice via random forests. *Bioinformatics* 32: 859-866.
924 Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J and Engelken J. 2015.
925 Hierarchical boosting: a machine-learning framework to detect and classify hard selective
926 sweeps in human populations. *Bioinformatics* 31: 3946-3952.
927 Quinlan JR. 1986. Induction of decision trees. *Machine Learning* 1: 81-106.
928 Raj A, Stephens M and Pritchard JK. 2014. fastSTRUCTURE: variational inference of
929 population structure in large SNP data sets. *Genetics* 197: 573-589.
930 Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P and Thornton KR. 2014. Landscape of
931 standing variation for tandem duplications in Drosophila yakuba and Drosophila
932 simulans. *Mol Biol Evol* 31: 1750-1766.
933 Ronen R, Udpa N, Halperin E and Bafna V. 2013. Learning natural selection from the site
934 frequency spectrum. *Genetics* 195: 181-193.
935 Rosenzweig BK, Pease JB, Besansky NJ and Hahn MW. 2016. Powerful methods for detecting
936 introgressed regions from population genomic data. *Mol Ecol* 25: 2387-2397.
937 Salazar C, Baxter SW, Pardo-Diaz C, Wu G, Surridge A, Linares M, Bermingham E and Jiggins
938 CD. 2010. Genetic evidence for hybrid trait speciation in Heliconius butterflies. *PLoS*
939 *Genet* 6: e1000930.

940 Salazar-Jaramillo L, Jalvingh KM, de Haan A, Kraaijeveld K, Buermans H and Wertheim B.
941 2017. Inter-and intra-species variation in genome-wide gene expression of Drosophila in
942 response to parasitoid wasp attack. *BMC Genomics* 18: 331.
943 Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N and Reich
944 D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*
945 507: 354-357.
946 Schrider DR, Houle D, Lynch M and Hahn MW. 2013. Rates and genomic consequences of
947 spontaneous mutational events in Drosophila melanogaster. *Genetics* 194: 937-954.
948 Schrider DR and Kern AD. 2016. S/HIC: Robust Identification of Soft and Hard Sweeps Using
949 Machine Learning. *PLoS Genet* 12: e1005928.
950 Schrider DR and Kern AD. 2017. Soft sweeps are the dominant mode of adaptation in the
951 human genome. *Mol Biol Evol* 34: 1863–1877.
952 Schrider DR, Shanku AG and Kern AD. 2016. Effects of Linked Selective Sweeps on
953 Demographic Inference and Model Selection. *Genetics* 204: 1207-1223.
954 Sheehan S and Song YS. 2016. Deep learning for population genetic inference. *PLoS Comput*
955 *Biol* 12: e1004845.
956 Shiao M-S, Chang J-M, Fan W-L, Lu M-YJ, Notredame C, Fang S, Kondo R and Li W-H. 2015.
957 Expression divergence of chemosensory genes between Drosophila sechellia and its
958 sibling species and its implications for host shift. *Genome Biol Evol* 7: 2843-2858.
959 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
960 polymorphism. *Genetics* 123: 585-595.
961 True JR, Weir BS and Laurie CC. 1996. A genome-wide survey of hybrid incompatibility factors
962 by the introgression of marked segments of Drosophila mauritiana chromosomes into
963 Drosophila simulans. *Genetics* 142: 819-837.
964 Turissini DA and Matute DR. 2017. Fine scale mapping of genomic introgressions within the
965 Drosophila yakuba clade. *bioRxiv*: 152421.
966 Turner TL, Hahn MW and Nuzhdin SV. 2005. Genomic islands of speciation in Anopheles
967 gambiae. *PLoS Biol* 3: e285.
968

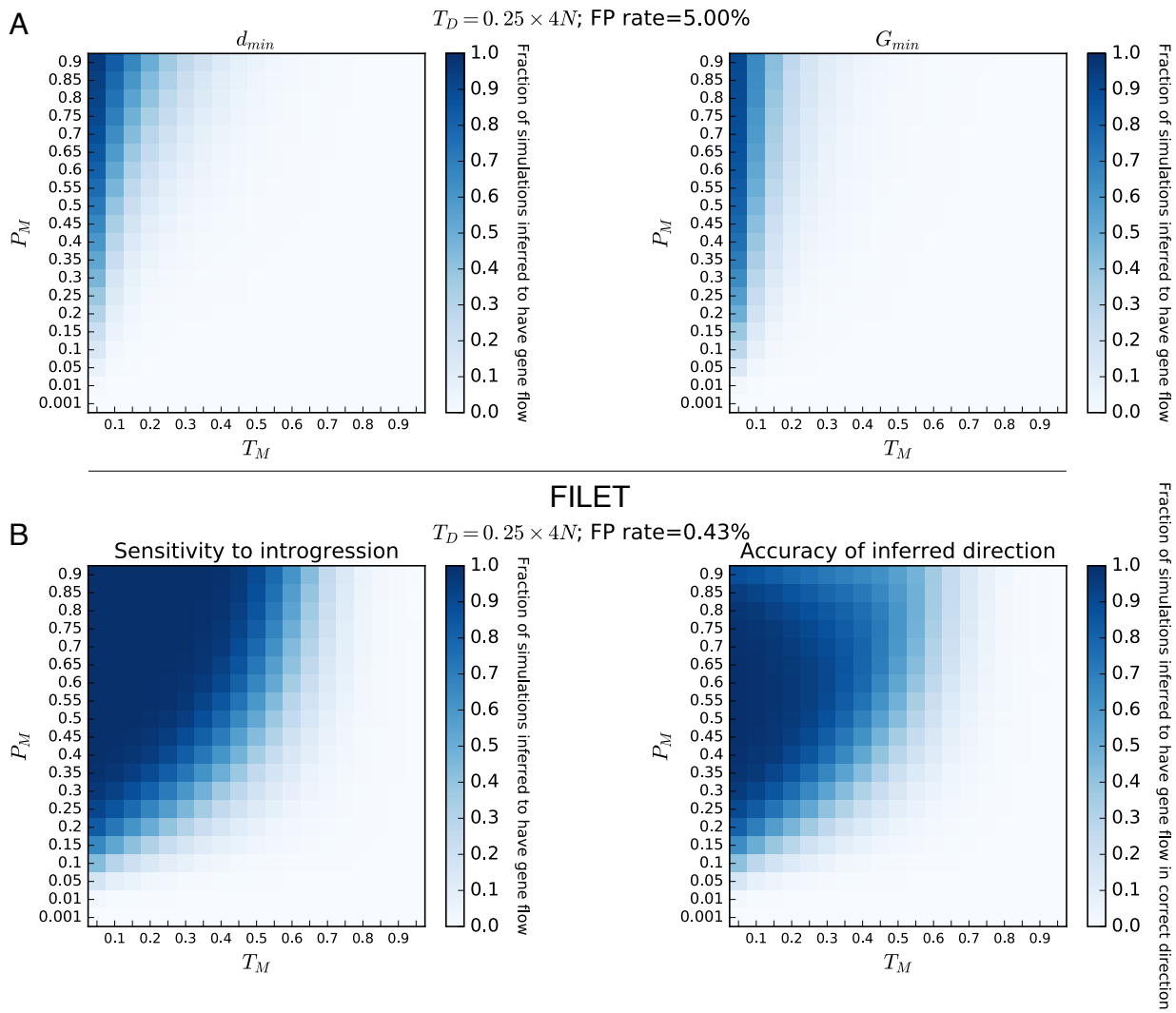

Fig. 1. Heatmaps showing several methods' sensitivity to detect introgression. We show the fraction of simulated genomic regions with introgression occurring under various combinations of migration times (T_M, shown as a fraction of the population divergence time T_D) and intensities (P_M, the probability that a given lineage will be included in the introgression event) that are detected successfully by each method. (*A*) Accuracy of d_{min} and G_{min} statistics, where a simulated region is classified as introgressed if the values of these statistics are found in the lower 5% tail of the distribution under complete isolation (from simulations). Thus, the false positive rate is fixed at 5%. (*B*) The accuracy of FILET on these same simulations. On the left we show the fraction of regions correctly classified as introgressed (compare to panel *A*). On the right, we show the fraction of all simulated regions that are not only classified as introgressed, but also for which the direction of gene flow was correctly inferred (i.e. if the direction is inferred with 100% accuracy for a given cell in the heatmap, the color shade of that cell will be identical to that in the heatmap on the left). The false positive rate, as determined from applying FILET to a simulated test set with no migration, is also shown.

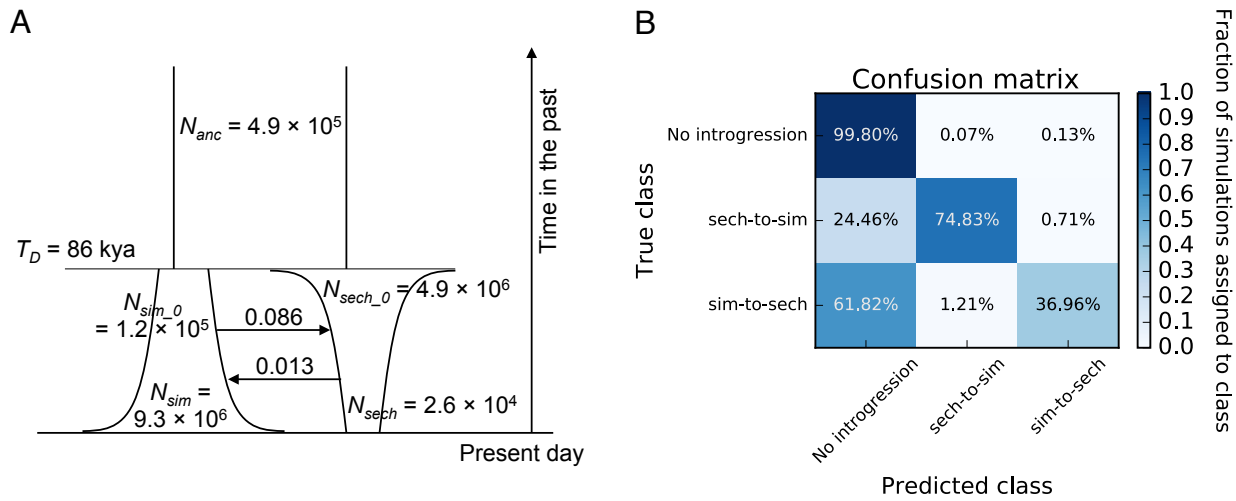

Fig. 2. Inferred joint population history of *D. simulans* and *D. sechellia*, and power to detect introgression under this model. (*A*) The parameterization of our best-fitting demographic model. Migration rates are shown by arrows, and are in units of $2 \times N_{anc} m$, where m is the probability of migration per individual in the source population per generation. (*B*) Confusion matrix showing FILET's classification accuracy under this model as assessed on an independent simulated test set. Perfect accuracy would be 100% along the entire diagonal from top-left to bottom-right, and the false positive rate is the sum of top-middle and top-right cells.

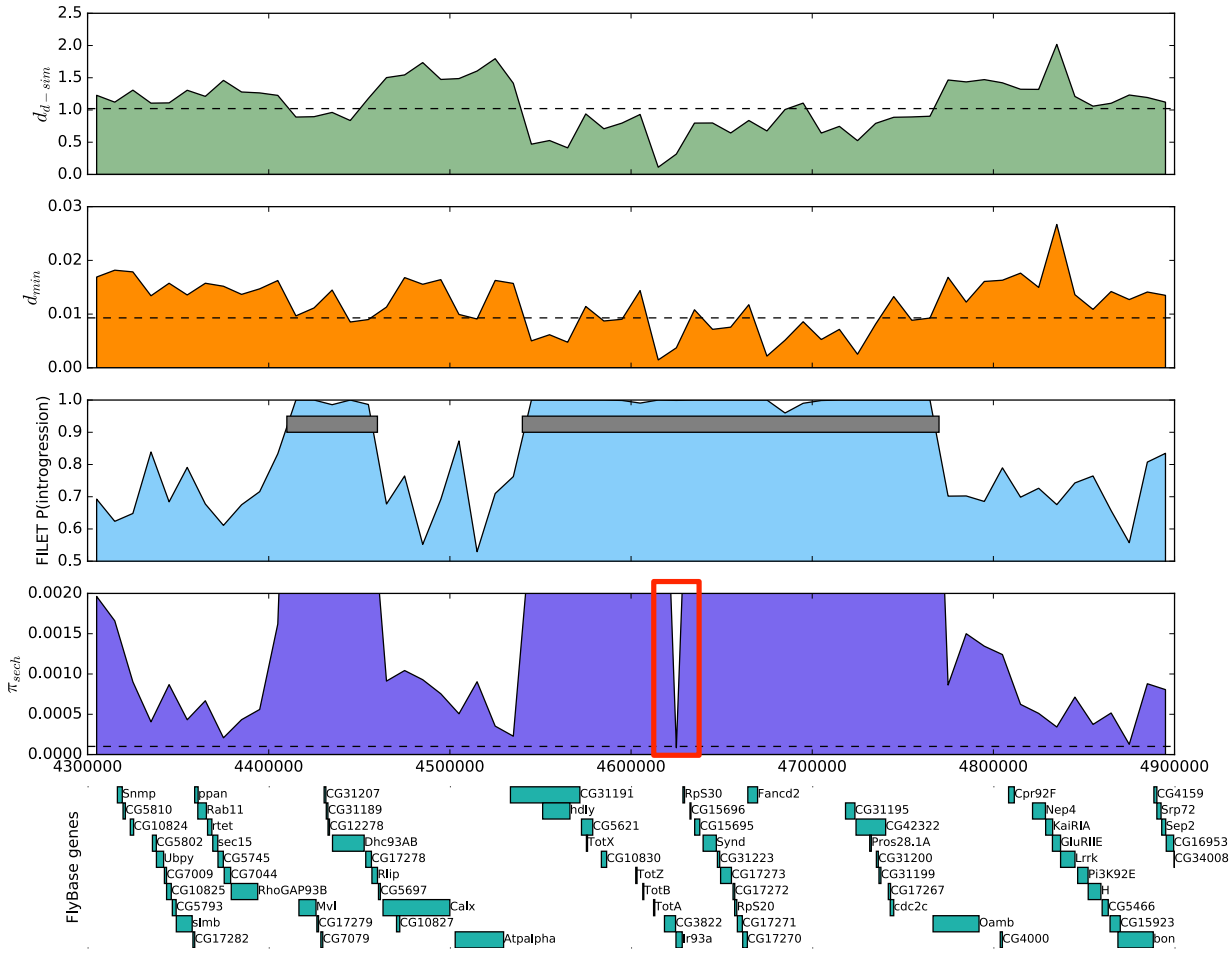

Fig. 3. A large genomic region on 3R classified by FILET as introgressed from *D. simulans* to *D. sechellia*. Values of the $d_{d\text{-}sim}$ and d_{min} (upper two panels) within each 10 kb window in the region are shown, along with the posterior probability of introgression from FILET (i.e. 1 – *P*(no introgression)). Clustered regions classified as introgressed are shown as gray rectangles superimposed over these probabilities. Also shown are windowed values of π in *D. sechellia*, with the sweep region highlighted in red, and the locations of annotated genes with associated FlyBase identifiers (Gramates et al. 2017).

1003 **SUPPLEMENTAL FIGURE AND TABLE LEGENDS**
1004
1005 **Figure S1.** Illustration of the difference in values of the d_{min} statistic calculated from joint
1006 population samples with and without introgression.
1007
1008 **Figure S2.** Violin plots showing the values of d_{min}, all four d_d statistics, and Z_X under simulated
1009 scenarios including introgression or lacking it for each values of T_D. The values of these statistics
1010 were obtained from the training data sets described in the Materials and Methods.
1011
1012 **Figure S3.** Heatmaps showing several methods' sensitivity to detect introgression. Same as
1013 Figure 1, but for other values of T_D. (*A*) Accuracy for d_{min} and G_{min} when $T_D = 1\times4N$ generations.
1014 (*B*) Accuracy of FILET when $T_D = 1\times4N$. (*C*) and (*D*) show the same when $T_D = 4\times4N$. (E) and
1015 (F) show the same when $16\times4N$.
1016
1017 **Figure S4.** Heatmaps showing FILET's sensitivity to introgression from an unsampled ghost
1018 population. (*A*) Sensitivity when $T_D = 0.25\times4N$ generations. (*B*) Sensitivity when $T_D = 1\times4N$
1019 generations. (*C*) $T_D = 4\times4N$ generations. (*D*) $T_D = 16\times4N$.
1020
1021 **Figure S5.** ROC curves showing power of FILET, d_{min} and G_{min} under each value of T_D. In order
1022 to generate these curves we transformed the classification task into a binary one: discriminating
1023 between isolation and introgression in either direction. (*A*) $T_D = 0.25\times4N$ generations. (*B*) $T_D =$
1024 $1\times4N$ generations. (*C*) $T_D = 4\times4N$. (*D*) $T_D = 16\times4N$. Training and test sets for these problems
1025 contained equal numbers of examples of introgression from population 1 into 2 and introgression
1026 from population 2 into 1.
1027
1028 **Figure S6.** ROC curves showing power of versions of FILET trained with decreasing numbers
1029 of training instances (ranging from 100 to 10000 for each class). (*A*) $T_D = 0.25\times4N$ generations.
1030 (*B*) $T_D = 1\times4N$ generations. (*C*) $T_D = 4\times4N$. (*D*) $T_D = 16\times4N$.
1031
1032 **Figure S7.** ROC curve showing FILET's power when trained and tested on simulated examples
1033 with different window sizes with $T_D = 0.25\times4N$ generations.
1034
1035 **Figure S8.** Population structure within *D. sechellia.* (*A*) The top three principal components of
1036 all *D. sechellia* diploid genomes. The cluster on the left shows the individuals from Praslin,
1037 while the cluster on the right shows all other individuals. Note that the cluster on the right is far
1038 less dispersed due to the very small amount of polymorphism among these individuals. The
1039 numbers in parentheses on each axis show the fraction of the variance explained by each
1040 principal component. (*B*) Results of running fastStructure on our *D. sechellia* samples with the
1041 number of subpopulations (*K*) ranging from 2 to 8.
1042

1043 **Figure S9.** Site frequency spectra of *D. sechellia* samples from Praslin, *D. sechellia* samples
1044 from all other locations, and *D. simulans* samples. The *D. sechellia* samples were both
1045 downsampled to n=12 as described in the text, while *D. simulans* was downsampled to n=18 (i.e.
1046 the same sample sizes used for our demographic inference). These SFS show the fraction of all
1047 polymorphisms found in each bin rather than the raw number of polymorphisms, and thus do not
1048 contain information about the total number of SNPs. As described in the text, there is >12-fold
1049 more polymorphism in the Praslin samples than in the non-Praslin samples.
1050
1051 **Figure S10.** Confusion matrix showing FILET's classification accuracy when trained under out
1052 inferred model of the *simulans-sechellia* joint demographic history, but applied to test data
1053 generated under a different model (described in Materials and Methods and shown in Table S4).
1054 under this model as assessed on an independent simulated test set. Perfect accuracy would be
1055 100% along the entire diagonal from top-left to bottom-right, and the false positive rate is the
1056 sum of top-middle and top-right cells.
1057
1058 **Table S1.** Results when applying FILET to simulations with constant bidirectional migration.
1059 10000 simulated replicates were tested for each parameter combination.
1060
1061 **Table S2.** Feature importance and rankings for each classifier used in this study.
1062
1063 **Table S3.** Sampling location, sequencing/mapping statistics, and SRA identifiers for each
1064 genome included in this study.
1065
1066 **Table S4.** Demographic parameter estimates inferred by ∂a∂i, along with a simple naïve model.
1067
1068